Neural Doubly Robust Proximal Causal Estimation

Ruolin Meng **Dhanajit Brahma** **Ricardo Henao** **Lawrence Carin**
Duke University

Abstract

We consider the challenging task of estimating treatment effects from observational data under the assumption that there are unobserved confounders. We employ the *proximal causal estimation* framework that assumes access to control (proxy) measurements that contain information about unobserved confounders. We consider outcome and treatment bridges, which provide two distinct ways of estimating causal effects. We also consider a doubly-robust approach, based on combining the outcome and treatment bridges, which is robust in expectation to either (but not both) of the two bridge functions being misspecified. We present a new theoretical bound on the estimation accuracy of the treatment bridge and we analyze the variance of the doubly-robust estimator. We investigate the impact of autoencoder-based regularization through an ablation study, finding that simpler models sometimes outperform more complex variants. Comparisons with state-of-the-art methods on synthetic and real-world data demonstrate the advantages of our approach.

1 INTRODUCTION

1.1 Proximal Methods & Double Robustness

A central and often untenable assumption in causal analysis is that within the observed data all common causes of the treatment and outcome variables, namely *the confounders*, have been measured (are observed) with sufficient fidelity to allow for a valid statistical adjustment (Glass et al., 2013). This assumption is inherently untestable from the data alone and, in practice, is frequently violated (Pearl et al., 2016). The consequence of this violation is biased causal effect estimates, which can lead to flawed decision-making in high-stakes settings (Rubin, 1974; Rosenbaum, 1984).

In response to this fundamental challenge, *proximal causal estimation* has emerged as a formal framework for identifying causal effects in the presence of unmeasured confounding (Kuroki and Pearl, 2014; Miao et al., 2018). The core principle of proximal methods is to leverage a pair of measured "proxies" to de-bias confounded estimates: as shown in Figure 1, a *treatment* control Z and an *outcome* control W (Tchetgen et al., 2020). Under specific and well-defined assumptions, this framework allows for the (nonparametric) identification of a causal effect that would otherwise be impossible to determine (Miao et al., 2018; Xu et al., 2021b; Cui et al., 2024; Meng et al., 2025).

The original formulation of proximal causal estimation methods is based on solving complex integral equations that are often ill-posed and challenging to implement (Cui et al., 2024). The integration of deep learning is a natural and necessary evolution of this framework. Using deep neural networks to model intricate relationships between proxies, treatments, and outcomes, practitioners can extend the proximal framework to scenarios where linear statistical models would fail. This fusion has enabled the development of advanced methods that are more robust, flexible, and applicable to a wider range of real-world problems (Mastouri et al., 2021; Kallus et al., 2021; Xu et al., 2021b; Kompa et al., 2022; Wu et al., 2024; Meng et al., 2025).

Despite its notable advantages, the proximal framework has its own set of assumptions, which are often untestable in practice and must be supported by strong subject matter knowledge and domain expertise. Consequently, more reliable approaches for proximal causal estimation are needed. In this direction, *doubly robust estimation* approaches, which have previously been considered in classical causal inference (without unobserved confounders) (Bang and Robins, 2005; Funk et al., 2011; Chernozhukov et al., 2018), have been extended to proximal methods (Kallus et al., 2021; Ghassami et al., 2022; Cui et al., 2024). The main advantage of these approaches is that they allow

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

for the separate specification of treatment and outcome bridges (models) in a manner that it is possible to provide consistent causal effect estimates even if either of them is misspecified, as long as one of them is correctly specified.

1.2 Contributions

In this work, we propose a doubly-robust proximal causal estimation approach that uses neural networks for its outcome and treatment bridge models. More specifically, we make the following contributions.

- We extend the results of Meng et al. (2025) for the outcome bridge to the treatment bridge, to characterize its error when the key assumption that the treatment bridge is not affected by the outcome control is violated.
- We provide new insights about the variance of the doubly robust estimator, namely, conditions under which its variance is lower than that of the treatment or outcome bridges separately.
- We perform an ablation study to investigate the impact of different modeling choices and the utility of regularization when learning the bridge functions.
- Experiments with synthetic and real-world data and comparisons with state-of-the-art baselines demonstrate the utility of the proposed approach. Importantly, the results for the real-world data are compared with independent results from a clinical trial.

1.3 Related Work

There has been recent machine learning research on proximal methods for causal inference with unobserved confounders, but only based on the outcome bridge (Xu et al., 2021a; Meng et al., 2025). The treatment bridge has been introduced recently, along with the double robust properties of learning both outcome and treatment bridges (Cui et al., 2024). However, the latter work focused principally on theoretical results and therefore assumed, in their experiments, that the explicit form of both bridges was known *a priori*, which is unrealistic in most applications.

Most prior work on proximal methods employed existence conditions that were technical (Cui et al., 2024; Miao et al., 2018, 2024; Xu et al., 2021a), and difficult to map to the information that proxy measurements must contain about the unobserved confounder. In (Meng et al., 2025) the authors developed a foundational information-theoretic bound on the quality of the outcome-bridge fit based on the relationship of the proxy measurements to each other and to the unobserved confounder. We develop new theory here in this direction, but for the treatment bridge.

There has been interest in modeling the proxy measurements, coupled with causal analysis, to enhance statistical strength. This has been pursued with an autoencoder (AE) (Meng et al., 2025) and a variational autoencoder (Louizos et al., 2017; Rissanen and Marttinen, 2021). Here we consider coupling the learning of the bridges with AEs for the proxy data, and provide new insights on the behavior (*e.g.*, variance of predictions) of these models in a doubly robust setting. Furthermore, we perform an ablation study exploring various types of regularization.

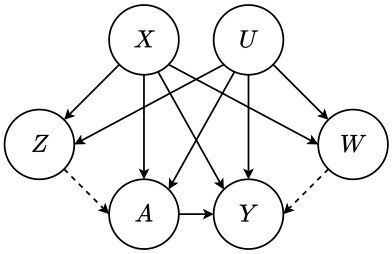

Figure 1: Graphical model of causal problem. Variables (X, U) are confounders, with X observed and U unobserved. The action (treatment) is A and the outcome Y. Z and W are proxy measurements. The dashed connections may or may not be present.

2 PROXIMAL BRIDGES

2.1 Assumptions

We consider causal estimation in the presence of observed and unobserved confounders X and U, respectively. The potential to account for U is made possible by also assuming access to proxy (control) measurements Z and W, with assumed properties developed below; see the graphical model for the setup in Figure 1. Proxies Z and W are denoted, respectively, as treatment and outcome control variables (Tchetgen et al., 2020); Z is a treatment control variable because the treatment (or action, A) may depend on Z in addition to dependence on (U, X), and W is an outcome control variable because the outcome Y may depend on W in addition to dependence on (U, X). While the theory is general, here we focus on a binary action $A \in \{0, 1\}$, and (continuous) outcome $Y \in \mathbb{R}$. The joint dependence of A and Y on (U, X) gives rise to confounding (Pearl et al., 2016; Rubin, 1974).

Consistent with prior work (Tchetgen et al., 2020; Xu et al., 2021b), we assume:

- $Y(a) \perp\!\!\!\perp A|U, X$ for $a \in \{0, 1\}$ *(latent ignorability)*;
- $W \perp\!\!\!\perp (A, Z)|U, X$ *(outcome control conditional independence)*;
- $Z \perp\!\!\!\perp Y|U, X, A$ *(treatment control conditional independence)*; and
- $0 < p(A = a|U, X) < 1$ almost surely for $a \in \{0, 1\}$ *(positivity)*.

This setting, often called *proximal causal estimation*, is discussed in Kuroki and Pearl (2014); Miao et al. (2018); Tchetgen et al. (2020); Miao et al. (2024).

2.2 Outcome Bridge

Theorem 1 (Miao et al. (2018)) *If there exists an outcome bridge function $h(W, X, a)$ that solves the integral equation*

$$\mathbb{E}[Y(a)|Z, X, A = a] = \mathbb{E}[h(W, X, a)|Z, X, A = a]\,, \quad (1)$$

almost surely, then

$$\mathbb{E}[Y|do(A = a)] = \mathbb{E}[h(W, X, a)]\,. \quad (2)$$

Theorem 1 shows a way to estimate the causal effect of treatment A on the outcome Y, without having access to the unobserved confounder U, through an outcome bridge function $h(W, X, a)$, which depends only on observed quantities W, X and A. The proof of Theorem 1 is described in Miao et al. (2018), and requires technical assumptions beyond those considered above, such as completeness (Xu et al., 2021b).

2.3 Treatment Bridge

Theorem 2 (Cui et al. (2024)) *If there exists a treatment bridge function $q(Z, X, a)$ that solves the integral equation*

$$\frac{1}{p(A = a|W, X, a)} = \mathbb{E}[q(Z, X, a)|W, X, A = a]\,, \quad (3)$$

almost surely, then

$$\mathbb{E}[Y|do(A = a)] = \mathbb{E}[\mathbb{I}(A = a)q(Z, X, a)Y]\,, \quad (4)$$

where $\mathbb{I}(\cdot)$ is an indicator function.

Like Theorem 1, Theorem 2 shows a way to estimate the causal effect in (2), without having access to the unobserved confounder U, through a treatment bridge $q(Z, X, a)$, which depends only on observed quantities Z, X and A. Note that intuitively (4) is the average of outcomes Y weighted by $q(Z, X, a)$, which acts as an estimator of (the inverse) treatment propensity based on observed quantities and restricted to observations such that $A = a$. The proof of Theorem 2 is described in Cui et al. (2024) and also involves technical assumptions that are typically difficult to verify in practice.

2.4 Doubly-Robust (DR) Estimator

Theorem 3 (Cui et al. (2024)) *The function*

$$\begin{aligned}\varphi_{DR}(a, W, X, Y, Z) \quad (5)\\ = \mathbb{I}(A = a)q(Z, X, a)[Y - h(W, X, a)]\\ + h(W, X, a)\,,\end{aligned}$$

yields

$$\mathbb{E}[Y|do(A = a)] = \mathbb{E}[\varphi_{DR}(a, W, X, Y, Z)]\,, \quad (6)$$

if <u>either</u> $h(W, X, a)$ or $q(Z, X, a)$ are solutions to their respective integral equations.

The proof of Theorem 3 can be found in Cui et al. (2024), and it requires additional technical requirements detailed there.

3 INFORMATION IN PROXIES

As discussed above, multiple technical assumptions are needed to prove Theorems 1 and 2, and these leave the needed properties of the proxies (W, Z) relative to U somewhat opaque. We next consider the error in fitting (1) and (3) in the presence of *imperfect* proxy measurements, from which we derive bounds on bridge quality tied to mutual information between (U, W, Z).

3.1 Outcome Bridge Approximation Error

Consider a solution of the Fredholm integral equation in (1) conditioned on $A = a$ and $X = x$.

Theorem 4 (Meng et al. (2025)) *Assume that $\mathbb{E}[Y(a)|W, X, U]$ is C-Lipschitz in U and that U is almost surely supported on a bounded set with radius $\|U\| \leq R$, there exist an outcome bridge function $h(W, X, a)$ for which*

$$\begin{aligned}\mathbb{E}_{Z|x,a}\Big|\mathbb{E}_{W|Z,x,a}[h(W, x, a)] - \mathbb{E}[Y(a)|Z, x, a]\Big| \quad (7)\\ \leq CR\sqrt{2I(U; Z|W, x, a)}\,,\end{aligned}$$

where $|\mathbb{E}_{W|Z,x,a}[h(W, x, a)] - \mathbb{E}[Y(a)|Z, x, a]|$ is the absolute value of the error to the Fredholm integral equation in (1) conditioned on (Z, x, a).

The proof of Theorem 4, which is modified here from Meng et al. (2025) to account for observed confounders X, is presented in Appendix A.

We highlight key attributes of the proof because these will be employed within our model of the outcome bridge. Specifically, for $A = a$, $Z = z$ and $X = x$, in general

$$\mathbb{E}[Y(a)|z, x, a] = \int dW\ h_0(W, z, x, a)p(W|z, x, a)$$

$$h_0(W, z, x, a) = \int dU\ \mathbb{E}[Y(a)|U, W, x]p(U|W, z, x, a)\,.$$

Theorem 4 is based on consideration of the outcome bridge

$$h(W, x, a) = \int dU\ \mathbb{E}[Y(a)|U, W, x]p(U|W, x, a)\,. \quad (8)$$

As emphasized in Meng et al. (2025), this form of the bridge *does not* assume $p(U|W, x, a) = p(U|W, z, x, a)$, and it *does not* assume $h_0(W, z, x, a) = h(W, x, a)$. It makes the weaker assumption that the expectations of $h_0(W, z, x, a)$ and $h(W, x, a)$ wrt $p(W|z, x, a)$ are equal.

3.2 Treatment Bridge Approximation Error

We introduce a new bound on the solution of the Fredholm integral equation in (3) conditioned on $A = a$ and $X = x$.

Theorem 5 *Assume that $1/p(A = a|U, X)$ is C-Lipschitz in U and that U is almost surely supported on a bounded set with radius $\|U\| \leq R$, there exists a treatment bridge $q(Z, X, a)$ for which*

$$\mathbb{E}_{W|x,a}\left|\mathbb{E}_{Z|W,x,a}[q(Z,x,a)] - \frac{1}{p(A=a|W,x)}\right| \leq CR\sqrt{2I(U;W|Z,x,a)}\,. \tag{9}$$

where $|\mathbb{E}_{Z|W,x,a}[q(Z,x,a)] - \frac{1}{p(A=a|W,x)}|$ is the absolute value of the error in the fit to the Fredholm integral equation (3) conditioned on (W, x, a).

Theorem 5 is proven in Appendix C, and it is based on the following lemma (proven in Appendix B).

Lemma 6 *For the graphical model in Figure 1, with assumptions in Section 2, the following identity holds:*

$$\frac{1}{p(A=a|W,x)} = \int dZ\; q_0(W,Z,x,a)p(Z|W,x,a)$$
$$q_0(W,Z,x,a) = \int dU\; \frac{p(U|W,Z,x,a)}{p(A=a|U,x)}\,. \tag{10}$$

To remove the dependence of W in $q_0(W, Z, x, a)$, the following model is introduced for the treatment bridge:

$$q(Z,x,a) = \int dU\; \frac{p(U|Z,x,a)}{p(A=a|U,x)}\,. \tag{11}$$

Analogously to the above discussion about the outcome bridge, this *does not* assume $p(U|W, Z, x, a) = p(U|Z, x, a)$, and it *does not* assume $q_0(W, Z, x, a) = q(Z, x, a)$. It makes the weaker assumption that the expectations of $q_0(W, Z, x, a)$ and $q(Z, x, a)$ wrt $p(Z|w, x, a)$ are equal.

Theorem 4 from Meng et al. (2025) and our new Theorem 5 yield the *symmetric* sufficient conditions on the conditional mutual informations $I(U; Z|W, x, a)$ and $I(U; W|Z, x, a)$, each of which should be small for, respectively, (1) and (3) to be solved accurately. These results indicate that W and Z should each be strong proxies for U.

3.3 Illustration with SEM data

To provide a quantitative sense of the bound in Theorem 5, we consider the data generated by the following structural equation model (SEM):

$$W = U+\epsilon_W,\quad Z = U+\epsilon_Z,\quad A = \mathbb{I}\{\alpha U+\beta Z+\epsilon_A > 0\}\,,$$

where U, ϵ_W, ϵ_Z and ϵ_A are scalar random variables, drawn from zero-mean Gaussian distributions with variances σ_U^2, σ_W^2, σ_Z^2 and σ_A^2. For this setup

$$p(A=1|U=u) = \Phi\left(\frac{(\alpha+\beta)u}{\sqrt{\sigma_A^2+\beta^2\sigma_Z^2}}\right), \tag{12}$$

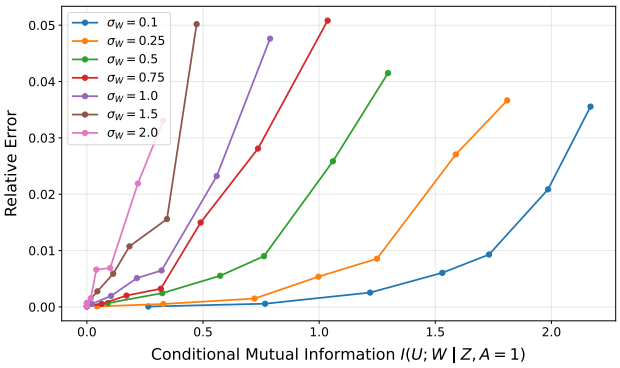

Figure 2: Relative approximation error *vs.* conditional mutual information averaged, *i.e.*, $\mathbb{E}_{W\sim p(W|A=a)}|\delta(W,a)|$ *vs.* $I(U; W|Z, A = 1)$, for the SEM data in Section 3.3. Each curve corresponds to a value of σ_W, and points on the curve correspond to $\sigma_Z = \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}$. For all points $\sigma_U = 10$, $\sigma_A = 10$, and $\alpha = \beta = 1$.

where $\Phi(\cdot)$ is its cumulative distribution function. We consider the relative error in the fit to (3):

$$\delta(W,a) = \frac{\mathbb{E}_{Z\sim p(Z|W,a)}[q_0(W,Z,a) - q(Z,a)]}{\mathbb{E}_{Z\sim p(Z|W,a)}[q_0(W,Z,a)]}\,. \tag{13}$$

Specifically, Figure 2 shows $\mathbb{E}_{W\sim p(W|A=a)}|\delta(W,a)|$, where all terms are computed based on the statistical properties of the underlying SEM generation process. On average, even when σ_Z and σ_W yield relatively noisy proxies (up to 20% of σ_U), the relative error is less than 5%.

3.4 Variance of the DR Estimator

The estimator $\mathbb{E}[\varphi_{DR}(a, W, X, Y, Z)]$ is unbiased if *either* the outcome $h(W, X, a)$ or the treatment $q(Z, X, a)$ bridge is correctly specified. We now consider analysis of the variance of the DR estimator, when *both* models are well specified.

Variance of DR Estimator relative to the variance of the outcome bridge Assume that the outcome bridge $h(W, X, a)$ is correctly specified, such that $\mathbb{E}[h(W, X, a)] = \mu(a)$, where $\mu(a)$ is the true causal effect under $do(A = a)$ and hence the DR estimator is also unbiased. The variance of the DR estimator, V_{DR}, can be expressed in relation to the variance of an outcome-model-only estimator, $V_O = \mathrm{Var}(h(W, X, a)) = \mathbb{E}\{[h(W, X, a) - \mu(a)]^2\}$:

$$V_{DR}(a) = V_O(a) + \mathbb{E}[1\{A=a\}q^2(Y-h)^2] - 2\mathbb{E}\Big[1\{A=a\}[h-\mu(a)]q[h-Y]\Big]\,,$$

where we write q for $q(Z, X, a)$ and h for $h(W, X, a)$ to simplify the form of the equation. Note that $\mathbb{E}[1\{A = a\}q^2(Y - h)^2] \geq 0$, and therefore for $V_{DR}(a) < V_O(a)$ the "cross term" $\mathbb{E}\Big[1\{A = a\}[h - \mu(a)]q[h - Y]\Big]$ must be sufficiently positive.

Variance of DR Estimator relative to the variance of the outcome bridge Now assume that the treatment model $q(Z, X, a)$ is correctly specified, such that $\mathbb{E}[\mathbb{I}(A = a)q(Z, X, a)Y] = \mu(a)$ and hence the DR estimator is also unbiased. Letting V_T be the variance of the treatment-only estimator, we have

$$V_{DR}(a) = V_T(a) + \mathbb{E}\Big[[h - 1\{A = a\}qh]^2\Big] - 2\mathbb{E}\Big[[1\{A = a\}qY - \mu(a)][1\{A = a\}qh - h]\Big]. \tag{14}$$

Since $\mathbb{E}\Big[[h - 1\{A = a\}qh]^2\Big] \geq 0$, for $V_{DR}(a) < V_T(a)$, there is a corresponding need for the cross term $\mathbb{E}\Big[[1\{A = a\}qY - \mu(a)][1\{A = a\}qh - h]\Big]$ to be sufficiently positive.

Lemma 7 *If the treatment bridge* and *outcome bridge are correctly specified, then*

$$\mathbb{E}\Big[[1\{A = a\}qY - \mu(a)][1\{A = a\}qh - h]\Big] = \mathbb{E}\Big[1\{A = a\}qYh[1\{A = a\}q - 1]\Big], \tag{15}$$

where $1\{A = a\}q[1\{A = a\}q - 1] \geq 0$ *for all* (X, Z)*, and*

$$\mathbb{E}\Big[1\{A = a\}[h - \mu(a)]q[h - Y]\Big] = \mathbb{E}\Big[[h - \mu(a)][h - \tilde{h}]\Big], \tag{16}$$

where

$$\tilde{h}(W, X, a) = \int dY dZ \; Y \; q(Z, a, X)p(Y, Z, A = a|W, X), \tag{17}$$

and $\mathbb{E}[\tilde{h}(W, X, a)] = \mathbb{E}[Y|do(A = a)]$.

The proof of Lemma 7 is provided in Appendix D.

Remark The form in (11) for the treatment bridge has the property that $q(Z, X, a) > 1$, since it is an expectation of the inverse probability $1/p(A = a|U, X)$. While the Fredholm integral equation in (3) may not have a unique solution, we proceed by assuming that the true solution is well-behaved and shares this property. This assumption guides both our practical model implementation and our subsequent theoretical analysis of the estimator's variance in Appendix D.

Since $1\{A = a\}q[1\{A = a\}q - 1] \geq 0$ for all (X, Z), the degree to which (15) is positive is guided by whether Y and $h(W, X, a)$ have the same sign, particularly where $q(Z, X, a)$ is large (it is at those points for which $A = a$ is improbable). Further, in relation to (16), large values of $q(Z, X, a)$ indicate that action $A = a$ is improbable. Through (17), for (X, Z) for which $A = a$ is improbable, it is reasonable to expect $\tilde{h}(W, X, a)$ to be distinct from $h(W, X, a)$; (16) will be positive if, in expectation wrt $p(X, W)$, the difference between h and $\tilde{h}$ is of the same sign as the difference between h relative to the overall mean $\mu(a)$ (of most interest when q is large).

It is difficult to test these conditions *a priori*, as they are dependent on the data generation process; however, this analysis suggests that it is possible for $V_{DR} < \min\{V_O, V_T\}$, and we will see this in our experiments.

3.5 Connection to Semiparametric Efficiency

The variance analysis in Section 3.4 can be understood in light of the classical semiparametric efficiency theory. We briefly review the relevant results and explain how the two perspectives relate.

In the semiparametric statistics literature, doubly robust estimators constructed from the efficient influence function (EIF) (Bickel et al., 1993; Tsiatis, 2006; Robins et al., 1994; Newey, 1994) are known to achieve the semiparametric efficiency bound, *i.e.*, the lowest asymptotic variance among all regular asymptotically linear estimators. The DR estimator φ_{DR} in (5) is precisely of this form: the centered function $\varphi_{\text{DR}} - \mu(a)$ is the EIF for the causal functional $\mu(a) = E[Y|do(A = a)]$ in the proximal model (Cui et al., 2024). When both the outcome bridge h and the treatment bridge q are consistently estimated at rates satisfying a product condition, specifically, the *product* of the estimation errors $\|\hat{h} - h_0\|$ and $\|\hat{q} - q_0\|$ must be $o_p(N^{-1/2})$, the DR estimator achieves $\sqrt{N}$-consistency and the semiparametric efficiency bound (Chernozhukov et al., 2018; van der Laan and Robins, 2003; Newey, 1990; van der Vaart, 2000). A consequence of this is that asymptotically $V_{\text{DR}} \leq \min(V_O, V_T)$. The use of flexible neural networks (which we introduce in the next section) for h and q is well-aligned with this framework, as recent work has established conditions under which neural estimators can achieve the convergence rates required for semiparametric inference (Farrell et al., 2021; Schmidt-Hieber, 2020; Chen and White, 1999; Bauer and Kohler, 2019).

While these asymptotic results provide important theoretical foundations, they require conditions beyond those needed for the analysis in Section 3.4. Both the semiparametric efficiency result and the analysis in Section 3.4 (specifically Lemma 7) assume that the bridges are correctly specified, and hence share the same underlying identification assumptions (including completeness). However, the semiparametric result additionally requires that the nuisance estimators $\hat{h}$ and $\hat{q}$ satisfy specific convergence rate conditions, and

that the sample size is large enough for the asymptotic approximation to be accurate. In practice, verifying that a particular neural network architecture achieves a specific convergence rate on a given problem is generally not feasible (Farrell et al., 2021), and at the sample sizes typical of applied work – including the experiments in this paper ($N = 2000$ for the SEM data, $N = 2901$ for the Framingham data, as detailed in Section 5) – whether the asymptotic regime has been reached is unknown.

The analysis in Section 3.4 does not depend on these additional conditions. Under the shared assumption of correct specification, Lemma 7 provides an exact decomposition of V_{DR} relative to V_O and V_T in terms of interpretable structural quantities. The cross term in (15) shows that variance reduction relative to V_T is governed by the behavior of Y and $h(W, X, a)$ in regions where $q(Z, X, a)$ is large – precisely where treatment $A = a$ is improbable and the reweighting by q has the greatest effect. The cross term in (16) shows that variance reduction relative to V_O depends on the relationship between $h(W, X, a)$ and $\tilde{h}(W, X, a)$ (defined in (17)), weighted by the deviation of h from the causal mean $\mu(a)$. These structural insights provide concrete, interpretable conditions under which variance reduction occurs, without requiring appeal to asymptotic rate theory.

The experiments in Section 5 are consistent with both perspectives. Figure 4 shows that when both bridges are well-specified (case (i)), the DR estimator achieves $V_{\mathrm{DR}} < \min(V_O, V_T)$ across all model variants, in agreement with both the semiparametric prediction and the favorable cross-term structure identified in Section 3.4. When one bridge is misspecified (cases (ii) and (iii)), the assumptions underlying both the efficiency result and Lemma 7 are violated; accordingly, the DR estimator retains consistency but variance reduction is no longer observed.

4 MODEL SPECIFICATION

4.1 Baseline Bridge Formulation (B)

Assume access to a dataset $\mathcal{D} = \{y_n, a_n, w_n, z_n, x_n\}_{n=1}^N$ of size N. We seek to learn outcome and treatment bridge functions, $h_\theta(W, X, a)$ and $q_\theta(Z, X, a)$, respectively, specified as neural networks, by optimizing them to satisfy their corresponding integral equations in (1) and (3). The form of our outcome bridge, motivated by (8), follows the construction in Meng et al. (2025), while the form of our model treatment bridge is motivated by (11). Specifically, we consider

$$h_\theta(W, x_n, a_n) = \mathbb{E}_{p(\epsilon)}\{\tilde{h}_\gamma[u_\phi^{(o)}(W, x_n, a_n, \epsilon), W, x_n, a_n]\}$$
$$q_\theta(Z, x_n, a_n) = \mathbb{E}_{p(\epsilon)}\{\tilde{q}_\gamma[u_\phi^{(t)}(Z, x_n, a_n, \epsilon), x_n, a_n]\},$$

where $p(\epsilon)$ represents a zero-mean isotropic Gaussian distribution, and $\theta = (\gamma, \phi)$. In order not to overload the notation, we have used θ to represent the parameters for both the outcome and treatment bridges, but these are, of course, distinct. These models are motivated by the idea that samples $u_\phi^{(o)}(W, x_n, a_n, \epsilon)$, with $\epsilon \sim p(\epsilon)$, represent (ideally) draws from $p(U|W, x_n, a_n)$. The model $u_\phi^{(t)}(Z, x_n, a_n, \epsilon)$ is similarly meant to represent draws from $p(U|Z, x_n, a_n)$. These models of the conditional distributions of the latent variables are motivated by the "reparameterization trick" originally developed in the context of the variational autoencoder (Kingma and Welling, 2013). We are motivated to consider these types of models, with $p(U|W, x_n, a_n)$ and $p(U|Z, x_n, a_n)$ modeled explicitly, because of our goal of coupling these models with autoencoders, as developed in Section 4.2.

Concerning the form of $\tilde{q}_\gamma(\cdot)$, because of the Remark in Section 3.4, we express this model as $\tilde{q}_\gamma(\cdot) = 1 + \exp[\psi_\gamma(\cdot)]$, where $\psi_\gamma(\cdot)$ is a neural network with real-valued output (details in Appendix E). This form of the treatment bridge was also considered in Cui et al. (2024), as discussed in Section 5.

The functions $u_\phi^{(o)}(\cdot)$ and $u_\phi^{(t)}(\cdot)$ are each represented by neural networks, as are $\tilde{h}_\gamma$ and $\tilde{q}_\gamma$. Details of all neural networks used in our experiments are provided in Appendix E.

When learning the bridge functions, we minimize:

$$\mathcal{L}_o(\theta, \phi) = \sum_{n=1}^N \left(y_n - \bar{h}_n\right)^2 \tag{18}$$

$$\mathcal{L}_t(\theta, \phi) = \sum_{n=1}^N \left(\frac{1}{p(A = a_n|w_n, x_n)} - \bar{q}_n\right)^2, \tag{19}$$

with

$$\bar{h}_n = \mathbb{E}_{p(W|z_n, x_n, a_n)}[h_\theta(W, x_n, a_n)] \tag{20}$$
$$\bar{q}_n = \mathbb{E}_{p(Z|w_n, x_n, a_n)}[q_\theta(Z, x_n, a_n)]. \tag{21}$$

In addition to the neural networks explicitly connected to the bridges, as discussed above, the expectations in $\bar{h}_n$ and $\bar{q}_n$ are calculated as averages over a set of samples obtained from $p(W|z_n, x_n, a_n)$ and $p(Z|w_n, x_n, a_n)$, respectively, which are modeled separately using generative models, *e.g.*, generative adversarial networks (GANs) Goodfellow et al. (2014) or diffusion models Ho et al. (2020). The details of these generative models depend on experimental considerations, and are discussed in Section 5 when presenting results. Note that we also model the propensity score $p(A = a|w, x)$ using standard methods (Austin, 2014; Li et al., 2018).

In practice, optimizing (18) or (19) involves four separate steps using $\mathcal{D}$, *i*) fit a propensity model $p(A = a|w, x)$; *ii*) fit generative models for $p(W|z_n, x_n, a_n)$ and $p(Z|w_n, x_n, a_n)$; *iii*) optimize the corresponding bridge function $h_\theta(W, X, a)$ and $q_\theta(Z, X, a)$; and *iv*) calculate ATE estimates using (2), (4) or (6).

Remark (On estimation of the propensity target). The loss $\mathcal{L}_t$ in (19) requires the target $1/p(A = a|W_n, X_n)$, which is itself estimated from data (here via logistic regression, as A is binary). The quality of the learned treatment bridge is therefore coupled to the quality of this propensity estimate. We intentionally adopt a simple approach to propensity estimation, as the methodological focus of this work is on the bridge-plus-autoencoder architecture and the doubly-robust framework. More sophisticated handling of this and other nuisance functions –for example, via sample splitting and cross-fitting in the spirit of debiased machine learning (Chernozhukov et al., 2018) – is a natural extension that could further improve the robustness of the treatment bridge estimation pipeline.

4.2 Enhancing Statistical Strength via AE

In developing the following autoencoder (AE) formulation, we do not consider covariates X, to simplify notation. Further, we develop the AE setup in the context of the treatment bridge, and an analogous setup is considered for the outcome bridge.

As discussed when introducing our outcome and treatment bridges, the model $u_\phi^{(t)}(Z, A, \epsilon)$ is meant to simulate draws from $p(U|A, Z)$, with ϵ a general random variable, where here we assume ϵ is isotropic Gaussian like in most prior work (Kingma and Welling, 2013; Louizos et al., 2017). We will leverage this model of $p(U|A, Z)$ within an AE model for (A, Z). Specifically, we will *share* $u_\phi^{(t)}(Z, A, \epsilon)$ within the treatment-bridge model and in a AE model for (A, Z), with the goal of enhancing statistical strength (and hence better learning $u_\phi^{(t)}(Z, A, \epsilon)$).

For the AE, we introduce additional models $p_\alpha(A|U, Z)$ and $p_\beta(Z|U)$ for A and Z, where (α, β) are their model parameters, respectively. We model $p_\alpha(A|U, Z)$ via a logistic regression framework, as A is binary, and we model $p_\beta(Z|U)$ as Gaussian. The details of these models as applied within our experiments are provided in Appendix E.

The AE loss may be expressed as

$$\mathcal{L}_{AE}(\phi, \alpha, \beta) = \sum_{i=1}^{N} \Big[\mathbb{E}_{p(U|z_n,a_n)} \log p_\alpha(a_n|U, z_n) + \mathbb{E}_{p(U|z_n,a_n)} \log p_\beta(z_n|U) \Big], \quad (22)$$

where, as discussed above, we use the "reparameterization trick" to represent the expectation wrt $p(U|z_n, a_n)$ in terms of samples from the reused $u_\phi^{(t)}(z_n, a_n, \epsilon)$.

When we learn a joint model, of the treatment bridge with the AE model (referred to as B+AE), we seek to *minimize* the total loss $\mathcal{L}_t - \lambda_{AE}\mathcal{L}_{AE}$, where $\lambda_{AE} > 0$ is a parameter, set via cross-validation.

When we present experiments in Section 5, we consider separate bridge-plus-AE setup for the outcome and treatment bridges, where the respective AEs model (A, Z) and (A, W). We also considered learning all of these together and imposing consistency in the statistics of U as modeled via the outcome and treatment bridges. The results from this more complicated setup were commensurate with modeling the outcome and treatment bridges separately, so we present the results with that simpler (separate) setup.

4.3 Entropy Regularization (H)

Our models are based on approximations to the distributions $p(U|A = a, Z = z)$ and $p(U|A = a, W = w)$, through models $u_\phi^{(o)}(W, x_n, a_n, \epsilon)$ and $u_\phi^{(t)}(Z, x_n, a_n, \epsilon)$. For the B and B+AE models referenced above, we have also considered the addition of an entropy-based regularization (Meister et al., 2020), referred to as B+H and B+AE+H. Such an entropy-based regularization encourages that the conditional distributions for U do not collapse to a point. We approximate the entropy using a differentiable kernel-based estimator over samples $\{u_i\}_{i=1}^M$. Following Kolchinsky and Tracey (2017), we use a Gaussian kernel to define:

$$\hat{H}(U) = \sum_{i=1}^{M} \log \Big(\frac{1}{M-1} \sum_{j \neq i} \exp \Big(- \frac{\|u_i - u_j\|^2}{2\gamma^2} \Big) \Big),$$

where γ is the kernel bandwidth, which is fixed or annealed during training. This estimator encourages diversity in the latent space by penalizing the concentration of u samples. Since each sample u_i constitutes a differentiable function through its corresponding encoder, the gradients of $\widehat{H}(U)$ are tractable via backpropagation. Entropy regularization has been widely used to improve model performance (Meister et al., 2020). We consider entropy regularization as part of our ablation study.

5 EXPERIMENTS

All models were developed using PyTorch, and each experiment was executed in a few minutes on a Tesla V100 PCIe 16 GB GPU. Code available at `https://github.com/ruolinmeng/NeuralDoublyRobustProximalCausalEstimation`.

Data We consider two datasets: the synthetic dataset originally used in Cui et al. (2024) based on

a SEM, and the real-world Framingham dataset (Benjamin et al., 1994) considered in Meng et al. (2025). The sizes of the two datasets are $N = 2000$ and $N = 2901$, respectively, and are split into 80% training, 20% validation. For the SEM data, the proxies W and Z were defined explicitly, while for the real-world dataset, W and Z were specified by "bucketing" observed covariates, in terms of which are most related to the outcome or treatment, as done in Meng et al. (2025) (we used the same definition of (W, Z) as in Meng et al. (2025)).

Models We compare the proposed method to *i*) deep feature proxy variable (DFPV) (Xu et al., 2021b); and *ii*) neural maximum moment restriction (NMMR) in its two variants, NMMR-U and NMMR-V (Kompa et al., 2022). Note that these are strong baselines to compare our method against, but they only provide estimators for the outcome bridge (do not allow comparisons for the treatment bridge or DR estimator). For the SEM data in Section 5, we also compare to the (linear) approach in Cui et al. (2024) based on M-estimators, which is appropriate for the special class of data considered there; the M-estimators are used to compute the outcome and treatment brides.

As an ablation study, we consider four variants of our method *i*) the basic bridge specification (B) in (18) and (19); *ii*) the autoencoder specification (B+AE) that uses (22); *iii*) the entropy-regularized approach (B+H) in Section 4.3 applied to B alone; and *iv*) the autoencoder with entropy regularization (B+AE+H). Models are optimized over the training set, model selection is based on the validation set loss, and ATEs are calculated on the training set (note that the ATE calculation is *distinct* from the original learning objective, of fitting Fredholm integral equations). Moreover, expectations over $p(W|z, x, a)$ and $p(Z|w, x, a)$ are calculated over 100 samples, whereas expectations over $p(\epsilon)$ use 10 samples. Details of the models and hyperparameter selection are presented in Appendix E.

Metrics We obtain the ATE using the following. Outcome bridge: $\psi_o = \mathbb{E}[h_\theta(W, X, 1)] - \mathbb{E}[h_\theta(W, X, 0)]$; Treatment bridge: $\psi_t = \mathbb{E}[(-1)^{1-A} q_\theta(Z, X, A)Y]$; and DR estimator: $\psi_{DR} = \mathbb{E}[(-1)^{1-A} q_\theta(Z, X, A)[Y - h_\theta(W, X, A)] + h_\theta(W, X, 1) - h_\theta(W, X, 0)]$.

For the synthetic data, we present the estimated median ATE with interquantile range (as a proxy for variance). For the real-world dataset (Framingham), as "ground truth" we use the ATE obtained from a separate randomized clinical trial using a Cox proportional hazard model (Yusuf et al., 2016). Results are summarized from 30 different seed runs. For the synthetic data, these include the data generation as well.

SEM Dataset The generative process for these data are as described in Cui et al. (2024). The data were designed such that the form of the outcome and treatment bridges was known *a priori*, and data were chosen as a proof-of-concept for the theory. In particular, the data were designed with

$$\begin{aligned} h_b(W, X, A) &= b_0 + b_a A + b_w + b_x X \\ q_t(Z, A, X) &= 1 + \exp\{(-1)^{1-A} t_0 + (-1)^{1-A} t_z Z \\ &\quad + (-1)^{1-A} t_a A + (-1)^{1-A} t_x X\}, \end{aligned}$$

where $b = (b_0, b_a, b_w, b_x)$ and $t = (t_0, t_z, t_a, t_x)$ are the *true* model parameters of the underlying bridges associated with the data generation. We will compare to the results in Cui et al. (2024), as a point of reference; however, this is an unfair comparison, as the M-estimator approach of Cui et al. (2024) assumes that the exact form of both bridges is known.

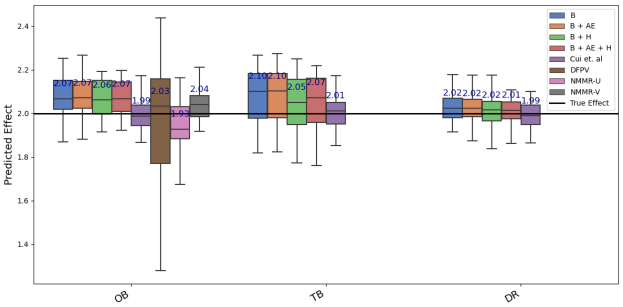

Figure 3: ATEs for the SEM dataset by different methods relative to the true ATE ($\psi = 2$). Results show median and IQRs as boxplots summarized over 30 different replicates.

The results in Figure 3 show the ATEs estimated by all baselines (Cui et al., 2024), DFPV, NMMR-U and NMMR-V), and variants of our model: B, B+H, B+AE and B+AE+H. We show separately estimates for the outcome bridge (OB), treatment bridge (TB), and double-robust (DR), but only show OB estimates for DFPV, NMMR-U, and NMMR-V, since they do not provide TB (and DR) estimates. We see that all methods result in comparable ATEs that are consistent with the ground truth ($\psi = 2.0$), however, we see that both the B+AE+H and the M-estimators of Cui et al. (2024) produce the closest (median) estimates.

By examining the results of our models, B, B+AE, B+H and B+AE+H (for which we *do not* assume *a priori* knowledge of the form of the bridges), the DR estimator produces the best median ATE estimator. Additionally, note that the variance of all variants of our DR estimators achieve $V_{DR} < \min(V_O, V_T)$, as was suggested to be possible in Section 3.4. Although all models perform similarly, the bias of the B+AE+H model is the smallest, as is its variance.

Effects of Bridge Misspecification We now reconsider the SEM data of Cui et al. (2024), with the goal of examining the capacity of the DR estimator to be effective if either the outcome or treatment bridges

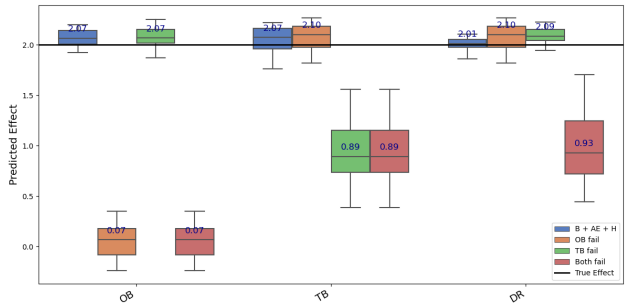

Figure 4: ATE for B+AE+H, compared to the true ATE ($\psi = 2$) when the outcome, treatment or both bridges are misspecified. Results show median and IQRs as boxplots summarized over 30 different replicates.

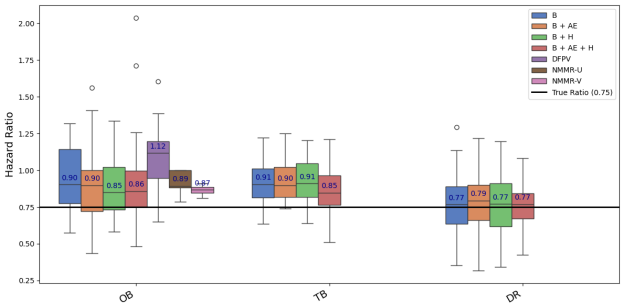

Figure 5: Hazard ratio (HR) by different methods compared to the true HR = 0.75. Results show median and IQRs as boxplots summarized over 30 different replicates.

are misspecified. In this experiment, we used the B+AE+H setup to learn the outcome and treatment bridges. However, to produce a poor (misspecified) model, we reduced the learning rate of our Adam optimizer to only 0.0001% of the learning rate used for the results in Figure 3, resulting in a poorly trained neural network. The results in Figure 4 show four cases: (*i*) both the outcome and treatment bridges learned well as in Figure 3, (*ii*) the outcome bridge is misspecified, (*iii*) the treatment bridge is misspecified, (*iv*) both models misspecified. The results in Figure 4 show robustness of the DR estimator in all cases, except (*iv*), as anticipated. Note that the DR estimator results in variance reduction under (*i*), but not under (*ii*) and (*iii*), as anticipated in Section 3.4.

Real-world Dataset We consider the Offspring cohort of the Framingham study Benjamin et al. (1994). To the authors' knowledge, this is the first examination of the proximal DR method with clinical-trial-based ground truth. We consider the ATE of statins (a cholesterol-lowering medication) to determine whether study participants will experience a cardiovascular event within 10 years. This means that both the treatment and the outcome are binary. Provided that in the original dataset participants may have been censored before the event horizon, they were excluded from the dataset. Additional details can be found in Appendix F. As ground-truth ATE, we use the hazard ratio reported from a separate large randomized control clinical trial, HR=0.75 (Yusuf et al., 2016). Unfortunately, we do not have access to individual estimates (for $A = \{0, 1\}$) to be able to calculate the ATE as in ψ_o, ψ_t and ψ_{DR}, however, we can still obtain estimates of the treatment and outcome bridge and the DR estimator and calculate the ratios, *e.g.*, $\text{HR}_{TB} = \mathbb{E}[h(W, X, A = 1)]/\mathbb{E}[h(W, X, A = 0)]$.

Figure 5 shows HR estimates for all models and estimators considered, including the baselines, among which NMMR-V is the best. In this experiment with real-world data, we see that the variance reduction in the DR estimator relative to OB and TB is less pronounced, likely due to model misspecification.

In Figure 5, while all DR estimators correctly identified a protective effect (HR < 1), the DR estimator built on the baseline bridge (B) in a DR setting achieved the lowest median bias. This highlights that for this specific dataset, the additional constraints imposed by the autoencoder and entropy regularizers did not lead to a more accurate estimate, thus it underscores the importance of model selection, as the optimal level of model complexity is often data dependent.

6 CONCLUSIONS

We have provided new theoretical results for proximal methods, showing how the information-theoretic quality of the proxy measurements impacts the treatment bridge, and we have investigated the variance properties of the double-robust estimator. We have also provided the first real-world demonstration of the utility of the DR setup, with validation through comparisons with a clinical-trial result.

Our ablation study of the components of the proposed model yielded important insights. While the autoencoder (AE) and entropy (H) regularizers are designed to enhance statistical strength by improving the latent-space representation of the proxies, we observed that they do not uniformly improve performance across all scenarios. On the real-world Framingham data, our most parsimonious doubly-robust model proved to be the most accurate in terms of bias (this may be due to limited data in this experiment). This finding suggests that, for some problems, a simpler and well-specified model may be sufficient and less prone to optimization challenges or overfitting than a more heavily regularized counterpart. Future work could focus on developing more formal guidelines or adaptive methods for choosing the optimal level of regularization in proximal causal models. More work is also needed on the design and selection of proxies (W, Z) in practice, building on Theorems 4 and 5.

Ackknowledgments

This work was supported by ONR grant number 313000130.

References

Peter C Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 2014.

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

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Emelia J Benjamin, Daniel Levy, Sonya M Vaziri, Ralph B D'Agostino, Albert J Belanger, and Philip A Wolf. Independent risk factors for atrial fibrillation in a population-based cohort: the Framingham heart study. *Jama*, 271(11):840–844, 1994.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993.

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

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

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

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, 2011.

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. In *International conference on artificial intelligence and statistics*, pages 7210–7239. PMLR, 2022.

Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34 (1):61–75, 2013.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

D.P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2013.

Artemy Kolchinsky and Brendan D Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Benjamin Kompa, David Bellamy, Tom Kolokotrones, Andrew Beam, et al. Deep learning methods for proximal inference via maximum moment restriction. *Advances in Neural Information Processing Systems*, 35:11189–11201, 2022.

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

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113 (521):390–400, 2018.

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*, 30:6446–6456, 2017.

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 *International conference on machine learning*, pages 7512–7523. PMLR, 2021.

C. Meister, E. Salesky, and R. Cotterell. Generalized entropy regularization or: There's nothing special about label smoothing. *Proc. Ass. Computational Linguistics*, 2020.

Ruolin Meng, Ming-Yu Ching, Dhanajit Brahma, Ricardo Henao, and Lawrence Carin. Coupling generative modeling and an autoencoder with the causal bridge. *arXiv preprint arXiv:2509.25599*, 2025.

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

Wang Miao, Xu Shi, Yilin Li, and Eric J Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects. *Statistical Theory and Related Fields*, 8(4):262–273, 2024.

Whitney K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Yury Polyanskiy and Yihong Wu. *Lecture Notes on Information Theory*. MIT, 2014.

Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. *Neural Information Processing Systems*, 2021.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Paul R Rosenbaum. From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, 79(385):41–48, 1984.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Yong Wu, Yanwei Fu, Shouyan Wang, and Xinwei Sun. Doubly robust proximal causal learning for continuous treatments. In *The Twelfth International Conference on Learning Representations*, 2024.

Anant Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, 2017.

L. Xu, H. Kanagawa, and A. Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems 34*, 2021a.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021b.

Salim Yusuf, Jackie Bosch, Gilles Dagenais, Jun Zhu, Denis Xavier, Lisheng Liu, Prem Pais, Patricio López-Jaramillo, Lawrence A Leiter, Antonio Dans, et al. Cholesterol lowering in intermediate-risk persons without cardiovascular disease. *New England Journal of Medicine*, 374(21):2021–2031, 2016.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix to Neural Doubly Robust Proximal Causal Estimation

A Outcome Bridge Details

A.1 Underlying assumptions connected to outcome bridge

For the outcome bridge $h(W, X, a)$, we seek to fit the following integral equation

$$\mathbb{E}[Y(a)|Z, X, a] = \mathbb{E}[h(W, X, a)|Z, X, a]\,. \tag{23}$$

Note the following identity

$$\mathbb{E}[Y(a)|Z, X, a] = \int dW\ \underbrace{\mathbb{E}[Y(a)|W, Z, X, a]}_{h_0(W,Z,X,a)} p(W|Z, X, a) \tag{24}$$

$$= \mathbb{E}[h_0(W, Z, X, a)|Z, X, a] \tag{25}$$

with

$$h_0(W, Z, X, a) = \int dU\ \mathbb{E}[Y(a)|W, U, X, a]p(U|W, X, Z, a) \tag{26}$$

where we have used $Y \perp\!\!\!\perp Z|U, X$, and therefore $\mathbb{E}[Y(a)|W, U, X, Z, a] = \mathbb{E}[Y(a)|W, U, X, a]$.

Importantly, $h_0(W, Z, X, a)$ is in general a function of (W, Z, X), while the desired bridge $h(W, X, a)$ is *not* a function of Z. We there postulate the following form for the outcome bridge

$$h(W, X, a) = \int dU\ \mathbb{E}[Y(a)|W, U, X, a]p(U|W, X, a)\,. \tag{27}$$

In general, $p(U|W, X, a) \neq p(U|W, X, Z, a)$, and $h(W, X, a) \neq h_0(W, Z, X, a)$. We require the much less strict requirement of

$$\mathbb{E}_{p(W|Z,X,a)}[h_0(W, Z, X, a)] = \mathbb{E}_{p(W|Z,X,a)}[h(W, X, a)]\,. \tag{28}$$

i.e., h_0 and h are equal *in conditional expectation* wrt $p(W|Z, X, a)$.

A.2 Proof of Theorem 4

We first consider the case $Z = z$ and $X = x$, and then we will average over (x, z). Fix $A = a$, $X = x$ and $Z = z$. For each $W = w$, define

$$\delta(w, x, z, a) := \mathbb{E}_{U\sim p(U|w,x,a)}[f(U)] - \mathbb{E}_{U\sim p(U|w,z,x,a)}[f(U)],$$

where $f(U) = \mathbb{E}[Y|a, x, w, U]$. Since f is C-Lipschitz and U is supported on a set of radius R, it follows from Pinsker's inequality and the standard variational form of KL divergence (see (Xu and Raginsky, 2017; Polyanskiy and Wu, 2014)) that:

$$|\delta(w, x, z, a)| = |h(w, x, a) - h_0(w, z, x, a)| \tag{29}$$

$$= |h(w, x, a) - \mathbb{E}[Y(a)|w, z, x, a]| \tag{30}$$

$$\leq CR \cdot \sqrt{2D_{\mathrm{KL}}(p(U|w, x, z, a)\,\|\,p(U|w, x, a))} \tag{31}$$

This inequality is related to a result in the proof of Lemma 1 in (Xu and Raginsky, 2017), using here that a C-Lipschitz function $f(U)$, with U supported on a ball of radius R, is CR-subGaussian (Vershynin, 2018).

We also note that

$$\begin{aligned}
\mathbb{E}_{Z|x,a}|\underbrace{\mathbb{E}_{W|Z,x,a}[h(W,x,a)] - \mathbb{E}[Y(a)|Z,x,a]}_{\text{error of Fredholm equation}}| &\leq \mathbb{E}_{W,Z|x,a}|h(W,x,a) - h_0(W,z,x,a)| && (32)\\
&\leq CR \cdot \mathbb{E}_{W,Z|x,a}\sqrt{2D_{KL}(p(U|W,x,Z,a)\|p(U|W,x,a)} && (33)\\
&\leq CR \cdot \sqrt{2\mathbb{E}_{W,Z|x,a}[D_{KL}(p(U|W,x,Z,a)\|p(U|W,x,a)]} && (34)
\end{aligned}$$

Consider that the conditional mutual information $I(U;Z|W,x,a)$ is defined as:

$$\begin{aligned}
I(U;Z|W,x,a) &= \int\int\int p(u,z,w|x,a)\log\left[\frac{p(u,z|w,x,a)}{p(u|w,x,a)p(z|w,x,a)}\right]du\,dz\,dw && (35)\\
&= \int\int\int p(u,z,w|x,a)\log\left[\frac{p(u|w,z,x,a)}{p(u|w,x,a)}\right]du\,dz\,dw && (36)
\end{aligned}$$

This can be rewritten as:

$$I(U;Z|W,x,a) = \int\int p(w,z|x,a)\left[\int p(u|w,z,x,a)\log\left[\frac{p(u|w,z,x,a)}{p(u|w,x,a)}\right]du\right]dw\,dz$$

The inner integral is precisely the KL divergence $D_{KL}(p(U|W,Z,x,a)\|p(U|W,x,a))$, so:

$$\begin{aligned}
I(U;Z|W,x,a) &= \int\int p(W,Z|x,a)\,D_{KL}(p(U|W,Z,x,a)\|p(U|W,x,a))\,dW\,dZ && (37)\\
&= \mathbb{E}_{p(W,Z|x,a)}[D_{KL}(p(U|W,Z,x,a)\|p(U|W,x,a))] && (38)
\end{aligned}$$

Consequently, we have

$$\mathbb{E}_{Z|x,a}|\mathbb{E}_{W|x,Z,a}[h(W,x,a)] - \mathbb{E}[Y(a)|Z,x,a]| \leq CR \cdot \sqrt{2I(U;Z|W,x,a)} \quad (39)$$

This theorem is a sufficient condition, in that it utilizes the postulated form of the outcome bridge in (27), but it may not be a necessary condition for the outcome bridge.

B Proof of Lemma 6

We first prove the relationship

$$\frac{1}{p(A=a|W,X)} = \int \frac{p(U|W,X,A=a)}{p(A=a|U,X)}dU\,. \quad (40)$$

For that, consider the following:

$$p(U|W,X,A=a) = \frac{p(A=a|U,W,X)p(U|W,X)}{p(A=a|W,X)}\,. \quad (41)$$

Rearranging terms we have

$$\frac{p(U|W,X)}{p(A=a|W,X)} = \frac{p(U|W,X,A=a)}{p(A=a|U,W,X)} = \frac{p(U|W,X,A=a)}{p(A=a|U,X)}\,, \quad (42)$$

where we have used the assumption $A \perp\!\!\!\perp W|(U,X)$. Integrating both sides of this equation wrt U, we arrive at (40).

Now we decompose the identity in (40) as

$$\frac{1}{p(A=a|W,X)} = \int \frac{p(U|W,X,A=a)}{p(A=a|U,X)} dU \tag{43}$$

$$= \int \Big[\int \frac{p(U|W,Z,X,A=a)}{p(A=a|U,X)} dU \Big] p(Z|W,X,A=a) dZ \tag{44}$$

$$= \int q_0(W,Z,X,a) p(Z|W,X,A=a) dZ \,. \tag{45}$$

with

$$q_0(W,Z,X,a) = \int \frac{p(U|W,Z,X,A=a)}{p(A=a|U,X)} dU \,. \tag{46}$$

C Treatment Bridge Details

C.1 Underlying assumptions connected to treatment bridge

Analogous to the outcome bridge, Lemma 6 suggests a postulated treatment bridge

$$q(Z,X,a) = \int \frac{p(U|Z,X,A=a)}{p(A=a|U,X)} dU \,. \tag{47}$$

The assumed treatment-bridge form *does not* assume $p(U|Z,X,a) = p(U|Z,X,W,a)$, and it *does not* assume $q_0 = q$. Rather, it makes the weaker assumption

$$\int dZ q_0(w,Z,x,a) p(Z|W=w,A=a,X=x) = \int dZ q(Z,x,a) p(Z|W=w,A=a,X=x) \,. \tag{48}$$

C.2 Proof of Theorem 5

Analogous to the above proof connected to the outcome bridge, we now consider

$$\delta(w,x,z,a) := \mathbb{E}_{U\sim p(U|z,x,a)}[f(U)] - \mathbb{E}_{U\sim p(U|w,z,x,a)}[f(U)],$$

where $f(U) = 1/p(A=a|U,x)$. Since f is C-Lipschitz and U is supported on a set of radius R, it follows from Pinsker's inequality and the standard variational form of KL divergence that:

$$|\delta(w,x,z,a)| = |q(z,x,a) - q_0(w,z,x,a)| \tag{49}$$

$$\leq CR \cdot \sqrt{2 D_{\mathrm{KL}}(p(U|w,x,z,a) \,\|\, p(U|z,x,a))}. \tag{50}$$

Consider that the conditional mutual information $I(U;W|Z,x,a)$ is defined as:

$$I(U;W|Z,x,a) = \int\int\int p(u,z,w|x,a) \log \left[\frac{p(u,w|z,x,a)}{p(u|z,x,a)p(w|z,x,a)} \right] du\, dz\, dw \tag{51}$$

$$= \int\int\int p(u,z,w|x,a) \log \left[\frac{p(u|w,z,x,a)}{p(u|z,x,a)} \right] du\, dz\, dw \tag{52}$$

This can be rewritten as:

$$I(U;W|Z,x,a) = \int\int p(w,z|x,a) \left[\int p(u|w,z,x,a) \log \left[\frac{p(u|w,z,x,a)}{p(u|z,x,a)} \right] du \right] dw\, dz$$

The inner integral is precisely the KL divergence $D_{KL}(p(U|W,Z,x,a) \| p(U|Z,x,a))$, so:

$$I(U;W|Z,x,a) = \int\int p(W,Z|x,a)\, D_{KL}(p(U|W,Z,x,a) \| p(U|Z,x,a))\, dW\, dZ \tag{53}$$

$$= \mathbb{E}_{p(W,Z|x,a)}[D_{KL}(p(U|W,Z,x,a) \| p(U|Z,x,a))] \tag{54}$$

from which

$$\mathbb{E}_{W|x,a}\Big|\mathbb{E}_{Z|W,x,a}[q(Z,x,a)] - \frac{1}{p(A=a|W,x)}\Big| \leq CR \cdot \sqrt{2I(U;W|Z,x,a)} \tag{55}$$

This completes the proof. Note that there is at least one treatment bridge that meets this bound, given in (47), proving a sufficient condition for meeting the inequality; however, it is possible that a treatment bridge may exist that could do better than this.

As a technical note, the Lipschitz assumption on $1/p(A=a|U,X)$ in Theorem 5 requires the propensity score to be bounded away from 0 and 1. This is more restrictive than the corresponding assumption in Theorem 4, though both theorems primarily serve to elucidate the information-theoretic requirements on the proxies (W,Z) rather than provide practical computational bounds, as $I(U;W|Z,x,a)$ and $I(U;Z|W,x,a)$ cannot be evaluated when U is unobserved.

D Proof of Lemma 7

We first consider the cross term

$$\begin{aligned}
&\mathbb{E}\Big[1\{A=a\}qY - \mu(a)][1\{A=a\}qh - h]\Big] \\
&\quad = \mathbb{E}\Big[1\{A=a\}qYh[1\{A=a\}q - 1]\Big] - \mu(a)\mathbb{E}\Big[[1\{A=a\}qh - h]\Big]
\end{aligned} \tag{56}$$

The key to the proof is to show $\mathbb{E}[1\{A=a\}qh] = \mathbb{E}(h)$, which implies that the second term in (56) vanishes. This is demonstrated as follows:

$$\begin{aligned}
\mathbb{E}[1\{A=a\}qh] &= \int dWdZdXp(W,Z,X,A=a)q(Z,X,a)h(W,X,a) && (57)\\
&= p(A=a)\int dWdX\ p(W,X|A=a)h(W,X,a)\int dZq(Z,X,a)p(Z|W,X,A=a) && (58)\\
&= \int dWdX\ p(W,X,A=a)h(W,X,a)\Big[\frac{1}{p(A=a|W,X)}\Big] && (59)\\
&= \mathbb{E}[h(W,X,a)] && (60)
\end{aligned}$$

where the relationship $\int dZq(Z,X,a)p(Z|W,X,A=a) = 1/p(A=a|W,X)$ comes from the assumption that q solves (3). Since $\mathbb{E}[1\{A=a\}qh] = \mathbb{E}(h)$, we have

$$\mathbb{E}\Big[1\{A=a\}qY - \mu(a)][1\{A=a\}qh - h]\Big] = \mathbb{E}\Big[1\{A=a\}qYh[1\{A=a\}q - 1]\Big] \tag{61}$$

which corresponds to (15) in the statement of Lemma 7. Note that this cross term arouse under the assumption that h is correctly specified, and to achieve (60), we assumed q is properly specified. So to achieve (61) it is assumed both bridges are properly specified.

Moving to the second cross term, we wish to justify the definition of $\tilde{h}(W,x,a)$, which is function that acts like an outcome bridge, in that $\mathbb{E}[Y|do(A=a),x] = \mathbb{E}[\tilde{h}(W,x,a)]$. Consider the following sequence of equations:

$$\begin{aligned}
\mathbb{E}[Y|do(A=a)] &= \int dX \int dU\ \mathbb{E}[Y(a)|U,X]p(U,X) && (62)\\
&= \int dX \int dU\ \mathbb{E}[Y(a)|U,X]\frac{p(U,X,A=a)}{p(A=a|U,X)} && (63)\\
&= \int dX \int dU\ \mathbb{E}[Y(a)|U,X]\int dZ\ q(Z,a,X)p(z|U,A=a,X)p(U,X,A=a) && (64)\\
&= \int dX \int dU\ \mathbb{E}[Y(a)|U,X,A=a]\int dZ\ q(Z,a,X)p(z|U,A=a,X)p(U,X,A=a) && (65)\\
&= \int dX \int dU \int dY(a)Y(a)p(Y(a)|U,X,A=a)\int dZ\ q(Z,a,X)p(z|U,A=a,X)p(U,X,A=a)\\
&= \int dX \int dU\ \int dY(a)\ Y(a)\int dZ\ q(Z,a,X)p[Y(a),Z,U,a,X] && (66)\\
&= \int dX\ \int dY\ \int dZ\ Y\ q(Z,a,X)p[Y,Z,X|A=a]p(A=a) && (67)\\
&= \int dXp(X,A=a)\ \int dY\ \int dZ\ Y\ q(Z,a,X)p[Y,Z|X,A=a] && (68)\\
&= \int dWdXp(W,X,A=a)\ \int dY\ \int dZ\ Y\ q(Z,a,X)p[Y,Z|W,X,A=a] && (69)\\
&= \int dWdXp(W,X)\ \underbrace{\int dY\ \int dZ\ Y\ q(Z,a,X)p[Y,Z,A=a|W,X]}_{\tilde{h}(W,X,a)} && (70)
\end{aligned}$$

This provides the definition of $\tilde{h}$, as specified. The treatment bridge $q(Z,X,a \geq 1$, and it is large in value at those (Z,X,a) for which action $A=a$ is improbable; at values of (W,X,Z) for which $A=a$ is improbable, we expect $\tilde{h}$ to potential be substantially different from h (the learning of which does not treat improbable actions with enhanced weighting).

Returning to the second cross term, we have

$$\begin{aligned}
&\mathbb{E}\Big[1\{A=a\}[h(W,X,a)-\mu(a)]q(Z,X,a)[h(W,X,a)-Y]\Big]\\
&\quad= \int [h(W,X,a)-\mu(a)]q(Z,X,a)[h(W,X,a)-Y]p(W,Z,X,Y,A=a)dWdZdXdY && (71)\\
&\quad= \int [h(W,X,a)-\mu(a)]q(Z,X,a)h(W,X,a)p(W,Z,X,A=a)dWdZdX\\
&\quad- \int [h(W,X,a)-\mu(a)]q(Z,X,a)Yp(W,Z,X,Y,A=a)dWdZdXdY && (72)
\end{aligned}$$

Let's first consider

$$\begin{aligned}
&\int [h(W,X,a)-\mu(a)]q(Z,X,a)h(W,X,a)p(W,Z,X,A=a)dWdZdX\\
&= \int [h(W,X,a)-\mu(a)]q(Z,X,a)h(W,X,a)p(W,Z,X|A=a)p(A=a)dWdZdX && (73)\\
&= \int dWdXp(W,X|A=a)p(A=a)[h(W,X,a)-\mu(a)]h(W,X,a)\int dZq(Z,X,a)p(Z|W,X,A=a) && (74)\\
&= \int dWdXp(W,X|A=a)p(A=a)[h(W,X,a)-\mu(a)]h(W,X,a)\frac{1}{p(A=a|W,X)} && (75)\\
&= \int dWdXp(W,X)[h(W,X,a)-\mu(a)]h(W,X,a) && (76)\\
&= \mathbb{E}_{p(W,X)}\Big[[h(W,X,a)-\mu(a)]h(W,X,a)\Big] && (77)
\end{aligned}$$

Now consider

$$\int [h(W,X,a)-\mu(a)]q(Z,X,a)Yp(W,Z,X,Y,A=a)dWdZdXdY$$

$$= \int [h(W,X,a)-\mu(a)]q(Z,X,a)Yp(W,Z,X,Y|A=a)p(A=a)dWdZdXdY \tag{78}$$

$$= p(A=a)\int dWdXp(W,X|A=a)[h(W,X,a)-\mu(a)]\int dZdYq(Z,X,a)Yp(Z,Y|W,X,A=a)dZdY$$

$$= \int dWdXp(W,X)p(A=a|W,X)[h(W,X,a)-\mu(a)]\int dZdYq(Z,X,a)Yp(Z,Y|W,X,A=a)dZdY \tag{79}$$

$$= \int dWdXp(W,X)[h(W,X,a)-\mu(a)]\int dZdYq(Z,X,a)Yp(Z,Y,A=a|W,X)dZdY \tag{80}$$

$$= \int dWdXp(W,X)[h(W,X,a)-\mu(a)]\tilde{h}(W,X,a) \tag{81}$$

Combining these relationships,, we have

$$\mathbb{E}\Big[1\{A=a\}[h(W,X,a)-\mu(a)]q(Z,X,a)[h(W,X,a)-Y]\Big]$$
$$= \int dWdXp(W,X)[h(W,X,a)-\mu(a)][h(W,X,a)-\tilde{h}(W,X,a)] \tag{82}$$

which corresponds to (16) in the statement of Lemma 7. The steps needed to arrive at (82) were based on the assumption that q was properly specified. To *interpret* this cross term from the standpoint of possible variance reduction, we also assume that h is properly specified.

E Models

E.1 NN Structures for SEM experiment

Outcome bridge models

$p(W \mid Z, A, X)$
Input(z, a, x)
FC$(4, 32)$, ReLU
FC$(32, 64)$, ReLU
FC$(64, 64)$, ReLU
Mean: FC$(64, 1)$
Std: FC$(64, 1)$, Softplus

$\tilde{h}_\gamma(U, W, X, A)$
Input(u, w, x)
FC$(4, 32)$, ReLU
FC$(32, 32)$, ReLU
FC$(32, 1)$

$p(Z \mid U, X)$
Input(u, x)
FC$(3, 16)$, ReLU
FC$(16, 32)$, ReLU
FC$(32, 1)$

$u_\phi^{(o)}(W, X, A, \epsilon)$
Input(w, x, ϵ)
FC$(4, 32)$, ReLU
FC$(32, 32)$, ReLU
FC$(32, 1)$

$p(A \mid U, Z, X)$
Input(u, z, x)
FC$(4, 32)$, ReLU
FC$(32, 64)$, ReLU
FC$(64, 1)$, Sigmoid

Treatment bridge models

$p(Z \mid W, A, X)$
Input(w, a, x)
FC$(4, 32)$, ReLU
FC$(32, 64)$, ReLU
FC$(64, 64)$, ReLU
Mean: FC$(64, 1)$
Std: FC$(64, 1)$, Softplus

$p(A \mid W, X)$
Input(w, x)
FC$(3, 32)$, ReLU
FC$(32, 1)$, Sigmoid

$u_\phi^{(t)}(Z, X, A, \epsilon)$
Input(z, x, ϵ)
FC$(4, 32)$, ReLU
FC$(32, 16)$, ReLU
FC$(16, 1)$

$\tilde{q}_\gamma(U, X, A)$
Input(u, x)
FC$(3, 32)$, ReLU
FC$(32, 16)$, ReLU
FC$(16, 1)$, Softplus

$p(W \mid U, X)$
Input(u, x)
FC$(3, 32)$, ReLU
FC$(32, 32)$, ReLU
FC$(32, 1)$

$p(A \mid U, X, Z)$
Input(u, x)
FC$(4, 32)$, ReLU
FC$(32, 32)$, ReLU
FC$(32, 1)$, Sigmoid

E.2 NN Structures for Framingham experiment

Outcome bridge models

$p(W \mid Z, A)$
Embedding Layer
Input(z, a)
FC$(17, 64)$, Linear
Time Step Embedding Layers
FC$(64, 64)$, Linear
SiLU Activation
FC$(64, 64)$, Linear
Projection Layer
FC$(16, 64)$, Linear
DDPM Structure
FC$(64, 256)$, ReLU, Dropout(0.1)
FC$(256, 512)$, ReLU, Dropout(0.1)
FC$(512, 256)$, ReLU, Dropout(0.1)
FC$(256, 16)$, Linear

$u_\phi^{(o)}(W, A, \epsilon)$
Input(w, ϵ)
FC$(17, 32)$, ReLU
FC$(32, 16)$

$\tilde{h}_\gamma(U, W, A)$
Input(u, w)
FC$(24, 32)$, ReLU
FC$(32, 1)$, Sigmoid

$p(A \mid U, Z)$
Input(u, z)
FC$(24, 8)$, ReLU
FC$(8, 1)$, Sigmoid

$p(Z \mid U)$
Input(u)
FC$(16, 32)$, ReLU
FC$(32, 16)$

Treatment bridge models

$p(Z \mid W, A)$
Embedding Layer
Input(w, a)
FC$(17, 64)$, Linear
Time Step Embedding Layers
FC$(64, 64)$, Linear
SiLU Activation
FC$(64, 64)$, Linear
Projection Layer
FC$(16, 64)$, Linear
DDPM Structure
FC$(64, 256)$, ReLU, Dropout(0.1)
FC$(256, 512)$, ReLU, Dropout(0.1)
FC$(512, 256)$, ReLU, Dropout(0.1)
FC$(256, 16)$, Linear

$p(A \mid W)$
Input(w)
FC$(16, 1)$, Sigmoid

$\tilde{q}_\gamma(U, A)$
Input(u)
FC$(16, 4)$, ReLU
FC$(4, 1)$, Softplus

$p(A \mid U, Z)$
Input(u)
FC$(24, 32)$, ReLU
FC$(32, 1)$, Sigmoid

$u_\phi^{(t)}(Z, A, \epsilon)$
Input(z, ϵ)
FC$(17, 32)$, ReLU
FC$(32, 16)$

$p(W \mid U)$
Input(u)
FC$(16, 32)$, ReLU
FC$(32, 16)$

F Details on Framingham Dataset with Binary Outcome

The Offspring cohort of the Framingham Heart Study (Benjamin et al., 1994) provides longitudinal data suitable for time-to-event analyses. To apply binary classification techniques, we can transform the outcomes by selecting a fixed time horizon and defining binary outcomes as follows:

1. Choose a fixed follow-up period (10 years) to assess whether the study participant died, *i.e.*, the event "death" occurred.
2. Defining Cases and Controls:
   - **Cases**: Participants who died within the 10-year period.
   - **Controls**: Participants who lived throughout the 10-year period.
3. **Exclusion Criteria:** Exclude participants who were *censored* within the 10-year period.

Dataset sample size: the training set included 2387 observations, 230 cases and 2157 controls, and the validation set included 514 observations, 63 Cases and 451 Controls. All other details of the dataset including covariate processing and split (into Z and W) are consistent with the experiment in (Meng et al., 2025).