PROXIMAL CAUSAL LEARNING OF OPTIMAL INDIVIDUALIZED DOSE RULE

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning optimal individualized treatment regimes from observational data is of great importance in precision medicine. However, when treatment is continuous and unmeasured confounding is present, identifying such regimes remains a significant challenge. In this paper, we leverage the proximal learning framework to nonparametrically identify and estimate the optimal individualized dose rule in the presence of unmeasured confounders. Corresponding theoretical guarantees for the proposed estimators are also established, including the excess risk bound and optimal convergence rate. Simulation studies and a synthetic experiment confirm the practical utility of our framework.

1 INTRODUCTION

The increasing availability of large-scale observational data enables personalized decision-making across domains such as healthcare and public policy (Einav & Levin, 2014). This has catalyzed extensive research into statistical and machine learning methodologies for optimal decision making (Collins & Varmus, 2015; Kosorok & Laber, 2019; Tsiatis et al., 2019; Athey & Wager, 2021; Robins, 1986; 1994; 1997; Jiang et al., 2019; Kitagawa & Tetenov, 2018).

A wide range of methods has been developed to identify the optimal individualized treatment regime. For instance, Qian & Murphy (2011); Chakraborty & Moodie (2013); Moodie et al. (2014) proposed an indirect method called Q-learning, which is essentially a regression-based approach by fitting the Q-function (Watkins & Dayan, 1992) and then assigning the treatment to each individual to maximize the Q-function value. On the other hand, Zhao et al. (2012) proposed a direct method called outcome-weighted learning (O-learning), which aims to find the optimal policy by directly maximizing the value function, which sheds light on various extensions like Chen et al. (2016) for continuous treatment and Zhou et al. (2017) for residual-weighted learning.

Most existing works for learning optimal individualized treatment regime from observational data rely on the unconfoundedness assumption. However, it is dubious that all the confounders are exactly measured in actual observational studies; thus, it is of great interest to mitigate the bias caused by unmeasured confounding and find the optimal policy. Recently, the proximal causal inference framework has been proposed as a powerful approach for causal identification in the presence of unmeasured confounding (Miao et al., 2018; Tchetgen Tchetgen et al., 2024; Cui et al., 2024). This framework achieves identification under unmeasured confounding by correctly classifying the observed covariates L into three types: $Z \in \mathcal{Z}$ as the treatment-inducing confounding proxy, $W \in \mathcal{W}$ as the outcome-inducing confounding proxy, and $X \in \mathcal{X}$ as the common causes of treatment and outcome. Then the bridge functions estimated with respect to proxy variables enable identification and estimation under unmeasured confounding.

Previous work on identifying the optimal policy within the proximal causal framework primarily focused on the binary treatments (Qi et al., 2024; Shen & Cui, 2023; Zhang & Tchetgen, 2024; Gao et al., 2025). However, extending existing methods to continuous treatments poses significant computational challenges, as it requires solving a nonlinear optimization for each individual, which can be unstable and computationally intensive (Chen et al., 2016). The primary objective of this paper is to estimate the optimal individualized dose rule (IDR) under the proximal causal inference framework.

Our contributions are summarized as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1. We establish the identification results of the values for two restricted IDR classes and leverage outcome-weighted learning methods to estimate the optimal IDR.
2. We provide a difference-of-convex algorithm to efficiently solve the policy optimization problem.
3. We derive the excess risk bound for our proposed estimators, showing their consistency and optimal convergence rate.

The remainder of the paper is organized as follows. Section 2 serves as a warm-up by reviewing the O-learning method assuming the validity of the unconfoundedness assumption. Section 3 extends to the unmeasured confounding setting, leveraging the proximal causal inference framework and obtains our identification results of the optimal regime. Section 4 details estimation algorithms for our proposed identification approaches. Section 5 establishes theoretical guarantees for the estimators. Sections 6 and 7 present simulation studies and the synthetic demand experiment to demonstrate the superior performance of our methods. Section 8 concludes with discussion and future directions.

2 PRELIMINARIES

We begin by reviewing the established framework for determining an optimal IDR in settings where all confounders are observed. Let $A \in \mathcal{A}$ denote a continuous treatment variable representing the dosage level. Let $L \in \mathcal{L}$ be a vector of observed pre-treatment covariates, such as patient demographics and clinical history, and Y the clinical outcome of interest. We assume Y is bounded and positive, following the convention that higher values of Y represent superior clinical outcomes. For any dose $a \in \mathcal{A}$, let $Y(a)$ denote the potential outcome that would have been observed had the individual received dose a. The observed dataset consists of n independent and identically distributed (i.i.d.) samples $\{L_i, A_i, Y_i\}_{i=1}^n$.

To identify the optimal IDR from observational data, it is generally assumed that unconfoundedness holds: $Y(a) \perp\!\!\!\perp A \mid L$ for all $a \in \mathcal{A}$. An IDR is a measurable function $d : \mathcal{L} \to \mathcal{A}$ mapping covariates to a recommended dose, and the class of all such regimes is denoted by $\mathcal{D}$. The optimal regime d^* maximizes the value function $\mathbb{E}[Y(d(L))]$, which under unconfoundedness reduces to:

$$d^*(L) = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y \mid L, A = a].$$

For continuous treatments, the value function cannot be directly identified via standard inverse probability weighting because $\mathbb{P}(A = d(L)) = 0$ almost surely. To address this, Chen et al. (2016) introduced a kernel-smoothed value estimator. They proposed estimating the optimal IDR by maximizing the following objective function:

$$\hat{d} = \arg\max_{d \in \mathcal{D}} \mathbb{E}_n \left[\frac{Y \mathbb{I}(|A - d(L)| \leq \phi)}{2\phi p(A \mid L)} \right],$$

where $\phi > 0$ is a bandwidth parameter and $\mathbb{E}_n$ denotes the empirical expectation. This objective function provides a consistent estimator for the true value function and can be efficiently optimized using techniques analogous to O-learning (Zhao et al., 2012) with a surrogate loss function to find a consistent estimate of the optimal IDR.

3 CONTINUOUS DOSE FINDING WITH UNMEASURED CONFOUNDING

In practical applications, the unconfoundedness assumption that all variables affecting both treatment and outcome are observed is often violated. If there exist unobserved confounders U that influence both the treatment A and the outcome Y, standard methodologies for learning an optimal IDR, such as those detailed in Section 2, yield biased results and are no longer applicable. To address this challenge, we leverage the proximal causal inference framework, as introduced by Tchetgen Tchetgen et al. (2024), to learn the optimal IDR in the presence of unmeasured confounding.

Following the proximal causal inference framework, we assume that the set of observed covariates L can be partitioned into three classes $L = (X, W, Z)$. Here, $X \in \mathcal{X}$ represents covariates that may influence both A and Y. The variables $W \in \mathcal{W}$ serve as outcome-inducing proxies (or negative control outcomes), which are related to the treatment A only through the latent confounders U and covariates X. Similarly, $Z \in \mathcal{Z}$ acts as a treatment-inducing proxy (or a negative control treatment), which is assumed to affect the outcome Y only through (U, X, A). We assume access to an observational dataset of n independent and identically distributed samples, denoted by $\{A_i, Y_i, X_i, W_i, Z_i\}_{i=1}^n$.

3.1 BASIC ASSUMPTIONS

Our methodology is grounded in standard assumptions for proximal causal inference, which are essential for identifying and estimating the optimal IDR in the presence of unmeasured confounding.

Assumption 1 *We make the following structural conditions:*

(1) Consistency: $Y = Y(A, Z)$ *and* $W = W(A, Z)$*, almost surely.*

(2) Positivity: $p(a \mid U, X) > 0$ *for any* $U, X \in \mathcal{U}, \mathcal{X}$*, almost surely.*

(3) Latent Unconfoundedness: $(Z, A) \perp\!\!\!\perp (Y(a), W) \mid (U, X)$ *for all* $a \in \mathcal{A}$*.*

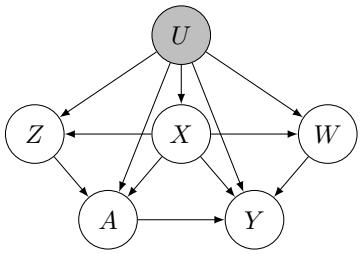


Figure 1: Causal DAG for the proximal causal framework.

Consistency and Positivity are general assumptions in the causal inference literature. Latent unconfoundedness implies that Z affects Y only through its influence on A, and that A and Z have no effect on W unless through (U, X). Figure 1 illustrates a classical setting corresponding to Assumption 1.

Assumption 2 *(Completeness of Outcome Proxy): For any* $a \in \mathcal{A}$*,* $x \in \mathcal{X}$*, and any square-integrable function* g*,* $\mathbb{E}[g(U) \mid W, A = a, X = x] = 0$ *almost surely if and only if* $g(U) = 0$ *almost surely.*

Assumption 3 *(Existence of Treatment Confounding Bridge): For any* $a \in \mathcal{A}$*, there exists a treatment confounding bridge function* $q_0(z, a, x)$ *that solves the following integral equation almost surely:*

$$\frac{1}{p(a \mid W, X)} = \mathbb{E}[q_0(Z, a, X) \mid W, A = a, X],$$

The globally optimal IDR d^*_{oracle} within the class of all policies mapping $(\mathcal{L}, \mathcal{U})$ to $\mathcal{A}$ is formally defined as:

$$d^*_{oracle}(L, U) = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Y \mid L, U, A = a].$$

Given that U is unobserved, d^*_{oracle} cannot be directly estimated from the data. The proximal framework established by Assumptions 1–3 enables the identification and estimation of restricted IDR classes using only observed variables.

3.2 OPTIMAL POLICY IDENTIFICATION WITH OUTCOME-INDUCING PROXIES

In this subsection, we propose a method to estimate the optimal IDR within a restricted policy class $\mathcal{D}_1$, comprising functions $d_1 : (\mathcal{X}, \mathcal{W}) \to \mathcal{A}_\epsilon$, where $\mathcal{A}_\epsilon = \{a \in \mathcal{A} : \text{dist}(a, \partial\mathcal{A}) \geq \epsilon\}$. The margin ϵ can be chosen small, so this restriction excludes only recommendations close to the boundary of the treatment support (Kennedy et al., 2017; Colangelo & Lee, 2026). We begin by establishing a fundamental identification theorem that serves as the theoretical basis for our optimization objective.

Theorem 1 *Assume that the conditional expectation of the potential outcome* $\mathbb{E}[Y(a) \mid U, X, W]$ *is continuous in* a*. Under Assumptions 1–3, the value function* $V(d_1) = \mathbb{E}[Y(d_1(X, W))]$ *for any IDR* $d_1 \in \mathcal{D}_1$ *is identified as follows:*

$$V(d_1) = \lim_{\phi \to 0^+} \mathbb{E}\left[\frac{Y\mathbb{I}(|A - d_1(X, W)| \leq \phi)}{2\phi} q_0(Z, A, X)\right].$$

Note that for $0 < \phi < \epsilon$, the symmetric kernel window around every recommended dose lies within $\mathcal{A}$, preserving unit mass. Building on this result, we define the smoothed value function:

$$\tilde{V}_\phi(d_1) = \mathbb{E}\left[\frac{Y\mathbb{I}(|A - d_1(X, W)| \leq \phi)}{2\phi} q_0(Z, A, X)\right].$$

As the bandwidth parameter ϕ vanishes, $\tilde{V}_\phi(d_1)$ converges to the true value $V(d_1)$. Consequently, the optimal IDR d_1^* can be approximated by $\tilde{d}_1^*$ which maximizes $\tilde{V}_\phi(d_1)$. This leads to the following conditional maximization problem:

$$\tilde{d}_1^*(X, W) = \arg\max_{a \in \mathcal{A}_\epsilon} \mathbb{E}\left[\frac{Y\mathbb{I}(|A - a| \leq \phi)}{2\phi} q_0(Z, A, X) \mid X, W\right].$$

This maximization problem is equivalent to minimizing the following risk functional, which can be interpreted as a weighted classification error:

$$\mathbb{E}\left[\frac{Y q_0(Z, A, X)}{2\phi}\right] - \tilde{V}_\phi(d_1) = \mathbb{E}\left[\frac{Y\mathbb{I}(|A - d_1(X, W)| > \phi)}{2\phi} q_0(Z, A, X)\right]$$

The indicator function $\mathbb{I}(|A - d_1(X, W)| > \phi)$ constitutes a zero-one loss, which is non-convex and discontinuous, posing significant computational challenges with empirical data (Zhang, 2004). To ensure computational tractability, we replace the zero-one loss with a continuous surrogate loss function. This leads to the following smoothed risk objective:

$$\mathscr{R}_\phi(d_1) = \mathbb{E}\left[\frac{Y\ell_\phi(A - d_1(X, W))}{\phi} q_0(Z, A, X)\right],$$

where $\ell_\phi(u) = \min(|u|/\phi, 1)$ represents a hinge loss serving as a surrogate for the zero-one loss.

We subsequently define the empirical risk $\hat{\mathscr{R}}_{\phi_n}(d_1)$, where the bandwidth ϕ_n is adaptively selected such that $\phi_n \to 0$ as $n \to \infty$:

$$\hat{\mathscr{R}}_{\phi_n}(d_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \hat{q}_0(Z_i, A_i, X_i)}{\phi_n} \ell_{\phi_n}(A_i - d_1(X_i, W_i)),$$

where $\hat{q}_0$ is a nonparametrically estimated treatment confounding bridge function. To prevent overfitting and control model complexity, we incorporate a regularization term. The final optimization problem is formulated as:

$$\min_{d_1 \in \mathcal{D}_1} \left\{\hat{\mathscr{R}}_{\phi_n}(d_1) + \lambda_n \|d_1\|_{\mathcal{D}_1}^2\right\},$$

where λ_n is a regularization parameter and $\|\cdot\|_{\mathcal{D}_1}$ denotes a norm on the function space $\mathcal{D}_1$. For instance, if $\mathcal{D}_1$ is a Reproducing Kernel Hilbert Space (RKHS), $\|\cdot\|_{\mathcal{D}_1}$ corresponds to the RKHS norm, facilitating the learning of complex and nonlinear decision rules.

3.3 Optimal policy identification with treatment-inducing proxies

In Section 3.2, we developed a framework for learning an optimal IDR $d_1(X, W)$ that relies on measured covariates X and the outcome-inducing proxy W. However, the treatment-inducing proxy Z also contains critical information regarding the latent confounder U. Intuitively, incorporating Z, or at least part of Z, into the decision rule could yield a more refined policy by capturing additional variation associated with unmeasured confounding. Consequently, we consider an expanded class of policies $\mathcal{D}_2$, where each IDR maps the full set of observed pre-treatment variables (X, W, Z) to a dose, i.e., $d_2 : \mathcal{X} \times \mathcal{W} \times \mathcal{Z} \to \mathcal{A}_\epsilon$.

We explore under what condition we can partially include Z into optimal regime construction. To achieve this, we assume that the treatment-inducing proxy Z can be decomposed into two distinct components. Without loss of generality, we denote the first component as Z, which may be directly relevant for policy-making, and the second component as Z^*, which serves as an auxiliary proxy and satisfies the following assumptions:

Assumption 4 *(Latent Unconfoundedness):* $(Z, Z^*, A) \perp\!\!\!\perp (Y(a), W) \mid (X, U)$ *for all* $a \in \mathcal{A}$.

Assumption 5 *(Existence of Auxiliary Treatment Confounding Bridge): For any* $a \in \mathcal{A}$*, there exists a treatment confounding bridge function* $q_1(z^*, a, x)$ *that satisfies the following integral equation almost surely:*

$$\mathbb{E}[q_1(Z^*, A, X) \mid W, A, X, Z] = \frac{1}{p(A \mid W, X, Z)},$$

Note that Assumptions 4 and 5 essentially require the observation of an additional valid treatment-inducing proxy Z^*, which is achievable as multiple proxies often exist in practical scenarios (Cui et al., 2024; Yu et al., 2025). We notice the distinction between Z and Z^* is not intrinsic and thus they are exchangeable. Furthermore, we assume the following completeness condition:

Assumption 6 *For any $a \in \mathcal{A}$, $x \in \mathcal{X}$, $z \in \mathcal{Z}$ and any square-integrable function g, $\mathbb{E}[g(U) \mid W, A = a, X = x, Z = z] = 0$ almost surely if and only if $g(U) = 0$ almost surely.*

Assumption 6 requires W to remain sufficiently informative about U after accounting for A, X, and Z. This introduces a trade-off between the number of decision-making variables and the required strength of the completeness condition for identification. Under this configuration, we establish the identification of the value function for policies in $\mathcal{D}_2$.

Theorem 2 *Assume that $\mathbb{E}[Y(a) \mid U, X, W, Z]$ is continuous in a and for every $a \in \mathcal{A}$, $p(a \mid U, X, Z) > 0$ almost surely. Under Assumptions 1, 4, 5 and 6, for any $d_2 \in \mathcal{D}_2$, the value function $V(d_2) = \mathbb{E}[Y(d_2(X, W, Z))]$ is identified as:*

$$V(d_2) = \lim_{\phi \to 0^+} \mathbb{E}\left[\frac{Y\mathbb{I}(|A - d_2(X, W, Z)| \leq \phi)}{2\phi} q_1(Z^*, A, X)\right].$$

Following the logic established in Section 3.2, we define a surrogate objective function using a continuous loss. The empirical risk using a pre-estimated auxiliary bridge function $\hat{q}_1$ is given by:

$$\hat{\mathscr{R}}_{\phi_n}(d_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \hat{q}_1(Z_i^*, A_i, X_i)}{\phi_n} \ell_{\phi_n}(A_i - d_2(X_i, W_i, Z_i)),$$

where $\ell_\phi(u) = \min(|u|/\phi, 1)$ denotes the surrogate loss. This leads to the final regularized empirical risk minimization problem for learning d_2:

$$\min_{d_2 \in \mathcal{D}_2} \left\{\hat{\mathscr{R}}_{\phi_n}(d_2) + \lambda_n \|d_2\|_{\mathcal{D}_2}^2\right\}, \tag{1}$$

where λ_n controls the penalty on model complexity measured by $\|\cdot\|_{\mathcal{D}_2}$.

4 Computational Algorithm

4.1 Estimation of confounding bridge functions

In this section, we detail the nonparametric estimation procedures for the two treatment confounding bridge functions q_0 and q_1, which were formally introduced in Assumptions 3 and 5, respectively.

Estimating the treatment bridges q_0 and q_1: To robustly estimate q_0, we leverage the minimax estimation framework with stabilizers, as proposed by Kallus et al. (2021). This framework reformulates the moment condition into the following minimax problem, which is particularly well-suited for nonparametric identification:

$$\begin{aligned}
\hat{q}_0 &= \arg\min_{q \in \mathcal{Q}} \max_{m \in \mathcal{M}} \mathbb{E}_n\left[\left(q(Z, A, X) - \frac{1}{p(A \mid W, X)}\right) m(W, A, X) - \lambda_m m^2(W, A, X)\right], \\
\hat{q}_1 &= \arg\min_{q \in \mathcal{Q}} \max_{m \in \mathcal{M}} \mathbb{E}_n\left[\left(q(Z^*, A, X) - \frac{1}{p(A \mid W, X, Z)}\right) m(W, A, X, Z) - \lambda_m m^2(W, A, X, Z)\right].
\end{aligned} \tag{2}$$

where $\mathcal{Q}$ and $\mathcal{M}$ denote the function classes for the estimator and the test function, respectively, such as RKHSs or neural networks, and λ_m is a stabilization parameter.

Estimating the generalized propensity score: Solving the optimization problems in Equation 2 requires estimating the inverse generalized propensity scores (GPS) $1/p(A \mid W, X)$ and $1/p(A \mid W, X, Z)$. Following Wu et al. (2023), we estimate them using conditional normalizing flows (CNFs) (Durkan et al., 2019; Rezende & Mohamed, 2015; Dinh et al., 2016; Sohn et al., 2015).

4.2 LEARNING ALGORITHMS FOR THE INDIVIDUALIZED DOSE RULES

We learn the optimal IDRs d_1 and d_2 by solving the regularized empirical risk minimization problems derived in the previous sections. We first consider a linear decision rule: $d_1(X, W) = \beta^T M^{(1)} + b$, where $M^{(1)} = (X^T, W^T)^T$ and the parameters to be estimated are $\Theta = (\beta^T, b)^T$. Define $w_i = Y_i \hat{q}_0(Z_i, A_i, X_i)$, $\mathcal{P} = \{i : w_i > 0\}$, and $\mathcal{N} = \{i : w_i < 0\}$, the corresponding empirical objective function $S(\Theta)$ is given by:

$$S(\Theta) = \frac{1}{n\phi_n} \sum_{i=1}^{n} w_i \min\left(\frac{|A_i - (\beta^\top M_i^{(1)} + b)|}{\phi_n}, 1 \right) + \frac{\lambda_n}{2} \|\beta\|_2^2. \tag{3}$$

Since the objective in Equation 3 is non-convex, we solve it by a Difference-of-Convex (DC) algorithm (An & Tao, 2005). Specifically, after writing $S(\Theta)$ as the difference of two convex functions $S_1(\Theta) - S_2(\Theta)$, the DC update is:

$$\Theta_{t+1} = \arg\min_{\Theta} \left(S_1(\Theta) - \left[S_2(\Theta_t) + \nabla S_2(\Theta_t)^T (\Theta - \Theta_t) \right] \right). \tag{4}$$

At iteration t, we define the residual $u_i^t = A_i - (\beta_t^\top M_i^{(1)} + b_t)$, its sign $s_i^t = \text{sign}(u_i^t)$ and the active set indicator $Q_i^{(t)} = \mathbb{I}(|u_i^t| \le \phi_n)$. Let $C_i = |w_i|/(n\lambda_n \phi_n^2)$ and define the linearization coefficient $h_i^{(t)} = C_i s_i^t \mathbb{I}\{i \in \mathcal{N} \text{ or } (i \in \mathcal{P},\ Q_i^{(t)} = 0)\}$. As detailed in Appendix C, each iteration reduces to solving the convex optimization problem:

$$\begin{aligned} \min_{\beta, b} \quad & \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^{n} h_i^{(t)} (\beta^\top M_i^{(1)} + b) + \sum_{i \in \mathcal{P}} C_i |A_i - (\beta^\top M_i^{(1)} + b)| \\ & + \sum_{i \in \mathcal{N}} C_i \big(|A_i - (\beta^\top M_i^{(1)} + b)| - \phi_n \big)_+ . \end{aligned} \tag{5}$$

The dual of Equation 5 is a Quadratic Program (QP):

$$\min_{\eta, v} \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \eta_i \eta_j \langle M_i^{(1)}, M_j^{(1)} \rangle - \sum_{i=1}^{n} \eta_i \left(A_i + \sum_{j=1}^{n} h_j^{(t)} \langle M_i^{(1)}, M_j^{(1)} \rangle \right) + \phi_n \sum_{i \in \mathcal{N}} v_i, \tag{6}$$

subject to:

$$\sum_{i=1}^{n} \eta_i = \sum_{i=1}^{n} h_i^{(t)}, \quad -C_i \le \eta_i \le C_i,\ \forall i, \quad v_i \ge \eta_i,\ v_i \ge -\eta_i,\ \forall i \in \mathcal{N}.$$

After solving this QP, the weight vector is updated as $\beta_{t+1} = \sum_{i=1}^{n} (\eta_i - h_i^{(t)}) M_i^{(1)}$, and b_{t+1} is obtained by minimizing the resulting one-dimensional convex primal objective. The algorithm iterates until the fitted decision rule converges.

To capture nonlinear relationships, we utilize the kernel trick. Assuming the RKHS $\mathcal{D}_1$ is equipped with a feature map Φ and kernel $K_1(\cdot, \cdot)$, the decision rule becomes $d_1(X, W) = \beta^T \Phi(M^{(1)}) + b$. Then the dual problem is obtained by replacing $\langle M_i^{(1)}, M_j^{(1)} \rangle$ in Equation 6 with $K_1(M_i^{(1)}, M_j^{(1)})$. When the final iteration ends, the learned nonlinear rule $d_1(X, W)$ is:

$$\sum_{i=1}^{n} \big(\eta_i - h_i^{(\text{final})} \big) K_1(M^{(1)}, M_i^{(1)}) + b_{\text{final}}. \tag{7}$$

In our experiments, we use the Gaussian kernel $K_1(x, x') = \exp(-\|x - x'\|_2^2 / \gamma_n^2)$.

The learning procedure for d_2^* is analogous, replacing $M^{(1)} = (X^T, W^T)^T$ with $M^{(2)} = (X^T, W^T, Z^T)^T$ and using the corresponding kernel K_2. The dual QP and final rule are unchanged.

5 THEORETICAL RESULTS

In this section, we establish the theoretical guarantees for our proposed estimators.We focus on $\hat{d}_2(X, W, Z)$, as analogous results hold for $\hat{d}_1(X, W)$. We first show the validity of our surrogate

approximation to the value function $V(d_2)$, then we derive a finite-sample excess risk bound for the regret of the estimated IDR and characterize the optimal convergence rate.

We begin by defining the surrogate value function $V_{\phi_n}(d_2)$ as the theoretical counterpart to the risk functional $\mathscr{R}_{\phi_n}(d_2)$:

$$V_{\phi_n}(d_2) := \mathbb{E}\left[\frac{Yq_1(Z^*, A, X)}{\phi_n}\right] - \mathscr{R}_{\phi_n}(d_2)$$
$$= \mathbb{E}\left\{\frac{Yq_1(Z^*, A, X)}{\phi_n} \max\left(1 - \frac{|A - d_2(X, W, Z)|}{\phi_n}, 0\right)\right\}.$$

We then quantify the approximation bias of this surrogate relative to the true value function $V(d_2)$.

Theorem 3 *For any measurable function $d_2 : (\mathcal{X}, \mathcal{W}, \mathcal{Z}) \to \mathcal{A}_\epsilon$, assume there exists a finite constant $C > 0$ such that the conditional expectation of the potential outcome is Lipschitz continuous in the dose, i.e., $\sup_{(a,U,X,W,Z)} |\frac{\partial}{\partial a}\mathbb{E}[Y(a) \mid U, X, W, Z]| \leq C$. When $0 < \phi_n < \epsilon$, it follows that:*

$$|V_{\phi_n}(d_2) - V(d_2)| \leq C\phi_n.$$

Theorem 3 shows that the surrogate bias is controlled by the bandwidth ϕ_n and vanishes as $\phi_n \to 0$. We next introduce regularity assumptions for analyzing the finite-sample performance of $\hat{d}_2$.

Assumption 7 *There exists a constant $C_1 > 0$ such that $\max\{|Y|, \|q_1\|_\infty\} \leq C_1$.*

Assumption 8 *The nuisance function estimator $\hat{q}_1$ satisfies $\|\hat{q}_1 - q_1\|_2 = O_p(n^{-\varphi})$ for some $\varphi > 0$. Furthermore, the estimator is bounded such that $\max\{\|\hat{q}_1\|_\infty\} \leq C_2$ for a constant $C_2 > 0$.*

Assumption 7 requires the outcome and the true treatment bridge function to be bounded, which is a standard condition in the literature (Zhao et al., 2019; Qi et al., 2024). Assumption 8 imposes conditions on the convergence rates of the nuisance function estimators. These rates are achievable by solving Equation 2 under appropriate regularity conditions (Kallus et al., 2021; Wu et al., 2023).

Under these conditions, we derive an excess-risk bound for $\hat{d}_2$, obtained by solving Equation 1 in the RKHS $\mathcal{D}_2$ equipped with a Gaussian kernel of bandwidth γ_n.

Theorem 4 *Assume that the optimal rule d_2^* over $\mathcal{D}_2$ belongs to $B^\alpha_{1,\infty}(\mathbb{R}^d)$ for some smoothness $\alpha > 0$, where $B^\alpha_{1,\infty}(\mathbb{R}^d) = \{f \in L_\infty(\mathbb{R}^d) : \sup_{t>0}(t^{-\alpha}\omega_{r,L_1(\mathbb{R}^d)}(f, t)) < \infty\}$ is a Besov space with ω the modulus of continuity, and assume that the distribution $P_{(X,W,Z)}$ has tail exponent $t \in (0, \infty]$. Under Assumptions 1–8, for any $\tau > 0$ and for choices of p satisfying $\frac{d}{d+t} < p < 1$, when $0 < \phi_n < \epsilon$, the following inequality for the value function holds with probability at least $1 - 3e^{-\tau}$:*

$$V(d_2^*) - V(\hat{d}_2) \leq \underbrace{c_1\left[\frac{1}{\gamma_n^{(1-p)(1+\delta)d}\lambda_n^p\phi_n^2 n}\right]^{\frac{1}{2-p}} + c_2\frac{\tau^{\frac{1}{2}}}{\phi_n n^{\frac{1}{2}}} + c_3\frac{\tau}{\phi_n n} + c_4\frac{\lambda_n}{\gamma_n^d} + c_5\frac{\gamma_n^\alpha}{\phi_n^2}}_{\text{I}} + \underbrace{c_6\frac{n^{-\varphi}}{\sqrt{\phi_n}}}_{\text{II}} + \underbrace{c_7\phi_n}_{\text{III}},$$

where $c_1, \ldots, c_7$ are constants independent of n.

Theorem 4 decomposes the excess risk bound into several components: the statistical error from finite sampling and the complexity of the RKHS $\mathcal{D}_2$ (I); the error from nuisance function estimation (II); and the bias from the kernel-smoothing approximation (III). By appropriately choosing the tuning parameters $\phi_n, \lambda_n, \gamma_n$, this bound ensures that $V(\hat{d}_2)$ converges to the optimal value $V(d_2^*)$. Assuming the nuisance convergence rate φ attains $1/4$ (Wu et al., 2023), denote $s = \min\left\{\frac{1}{6}, \frac{\alpha}{4\alpha+3[1+(1-p)\delta]d}\right\}$, we select the tuning parameters to balance the terms in the error bound as follows:

$$\lambda_n = n^{-(1+3d/\alpha)s}, \quad \gamma_n = n^{-3s/\alpha}, \quad \phi_n = n^{-s}.$$

This leads to the following optimal convergence rate:

$$V(d_2^*) - V(\hat{d}_2) = O_p\left(n^{-s}\right).$$

When the optimal rule d_2^* is sufficiently smooth ($\alpha \to \infty$), the optimal rate approaches $n^{-1/6}$, below the $n^{-1/4}$ typically observed in settings without unmeasured confounding (Chen et al., 2016), with the gap mainly constrained by the convergence rates of the nuisance function estimators.

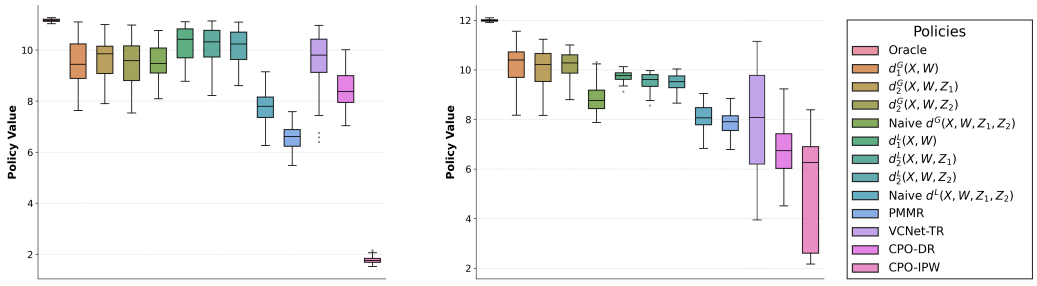
6 SIMULATION

To evaluate the performance of the proposed methods, we conduct simulation studies based on a data-generating process adapted from Mastouri et al. (2021) for (A, U, X, W, Z). We examine two outcome-generating scenarios. In Scenario 1, Y is generated under a linear oracle dose rule, whereas in Scenario 2, Y is generated under a nonlinear oracle dose rule. The complete data-generating mechanisms are provided in Appendix D.1.

To evaluate the proposed framework, we estimate the linear policies $d_1^L(X, W)$, $d_2^L(X, W, Z_1)$, and $d_2^L(X, W, Z_2)$, where the latter two use Z_2 and Z_1 as auxiliary treatment-inducing proxies, respectively, and their Gaussian-kernel counterparts defined in Equation 7. We first consider three baselines that assume no unmeasured confounding: (i) outcome-weighted learning under unconfoundedness (Chen et al., 2016), denoted by Naive $d^L(X, W, Z)$ and Naive $d^G(X, W, Z)$; (ii) continuous policy optimization (CPO; Kallus & Zhou, 2018), using self-normalized inverse-propensity weighting (CPO-IPW) or a doubly robust objective (CPO-DR) over clipped policies; and (iii) VCNet with functional targeted regularization (VCNet-TR; Nie et al., 2021), followed by maximization of the corrected outcome prediction. These baselines use (X, W, Z) under observed-covariate unconfoundedness. We additionally consider the regression-based baseline (iv) proxy maximum moment restriction (PMMR; Mastouri et al., 2021), which estimates the outcome bridge $h(A, W, X)$ using moments conditional on (A, Z, X) and selects the dose maximizing $\hat{h}(a, W, X)$. PMMR and VCNet-TR search 101 equally spaced doses in $[-5, 5]$. The oracle rules in Equation 17 and Equation 19, denoted by d_{oracle}, provide a benchmark.

Regarding the functional spaces, $\mathcal{D}_1$ and $\mathcal{D}_2$ are chosen as RKHSs equipped with Gaussian kernels. CNFs are used to estimate the generalized propensity score in both simulations and synthetic experiments, while the bridge functions q_0 and q_1 are estimated via Equation 2 in $\mathcal{Q}$ and $\mathcal{M}$. The baselines use the implementation settings detailed in Appendix D.2. For the DC algorithm, the tuning parameters ϕ_n, λ_n, and γ_n are determined via cross-validation.

For each estimated IDR, the value function is computed by substituting the predicted doses into the true outcome mechanisms and averaging the results over the test population. Figure 2 reports the results at $n = 1,500$; the corresponding $n = 3,000$ results are given in Appendix D.3. Our proposed methods achieve higher estimated policy values than both the regression-based baselines and methods that do not account for unmeasured confounding when the policy class is correctly specified: linear rules perform best in Scenario 1, while Gaussian-kernel rules perform best in Scenario 2. The same pattern holds at $n = 3,000$. We also observe that $d_2(X, W, Z_1)$, which utilizes Z_2 as the auxiliary treatment-inducing proxy, performs better than $d_2(X, W, Z_2)$. This is consistent with the fact that Z_2 serves as a more informative proxy for the latent confounder U (Mastouri et al., 2021).



(a) Scenario 1

(b) Scenario 2

Figure 2: Boxplots of estimated values at sample size $n = 1500$.

7 SYNTHETIC STUDY

We evaluate the proposed methods using the demand synthetic experiment, originally introduced by Hartford et al. (2017) for instrumental variable regression and later adapted to the proximal setting

by Xu et al. (2021). This data generating process is designed to study the effect of airline ticket price on sales. Our objective is to learn an individualized decision rule that maximizes sales Y. The continuous treatment A represents the airline ticket price, while the unmeasured confounder U captures latent demand. To mitigate unmeasured confounding, we employ fuel cost variables $Z = (Z_1, Z_2)$ as treatment-inducing proxies. The outcome-inducing proxy W is the number of views on a ticket reservation website. The structural equations are provided in Appendix E.

We assess performance using simulated datasets with sample sizes of 1,000, 2,000, and 4,000, each replicated over 100 independent trials. For interpretability, our proposed rules are linear: $d_1^L(W)$, $d_2^L(W, Z_1)$, and $d_2^L(W, Z_2)$, estimated as in Section 6. Among the baselines considered in the simulation study, we include only Naive $d^L(W, Z_1, Z_2)$, CPO-IPW, and CPO-DR here, as these methods can learn linear policies in our implementations, whereas PMMR and VCNet-TR do not. All three baselines use the full set of covariates (W, Z_1, Z_2) as predictors. All recommended prices are restricted to $[5, 50]$. We compute the value functions $V(\cdot)$ for both the proposed and competing approaches via the synthetic data generation process. As shown in Table 1, the proposed rules $d_1^L(W)$, $d_2^L(W, Z_1)$, and $d_2^L(W, Z_2)$ consistently achieve the highest estimated value.

Table 1: The value functions for different fitted rules across various sample sizes n.

Method	$n = 1000$	$n = 2000$	$n = 4000$
CPO-DR	48.5040 ± 3.9561	50.1646 ± 4.0524	49.4639 ± 3.6553
CPO-IPW	47.7841 ± 9.7110	46.9208 ± 7.8499	47.9852 ± 8.5883
Naive	46.1301 ± 0.4091	46.0052 ± 0.2709	45.9943 ± 0.2591
$d_1^L(W)$	62.2585 ± 4.0412	63.5741 ± 2.7905	63.5839 ± 2.8681
$d_2^L(W, Z_1)$	$\mathbf{64.0996 \pm 0.3890}$	64.0027 ± 0.3584	63.8481 ± 0.2899
$d_2^L(W, Z_2)$	63.2059 ± 3.2900	$\mathbf{64.2883 \pm 0.2726}$	$\mathbf{64.3318 \pm 0.2548}$

As a qualitative analysis, we further examine the fitted coefficients of the learned linear rules. We observe a clear distinction between the proposed proximal methods and the baselines. For the website-view proxy W, baseline rules such as CPO-DR,CPO-IPW and Naive yield a large positive coefficient, implying that higher views would dramatically increase the recommended price. In contrast, the proximal methods produce a coefficient that remains positive but is substantially smaller. This discrepancy arises because W reflects latent demand drivers U (e.g., holidays) that simultaneously influence historical prices and sales. Baselines that do not adjust for U mistakenly attribute demand-driven variation to W, thereby learning an overly aggressive pricing rule. The proximal methods, by adjusting for latent demand, yield a more moderate pricing response to website views. A similar pattern holds for the fuel-cost proxies: the coefficients under the proximal rules are relatively more moderate than those from the baselines, suggesting that the proposed methods are less prone to absorbing confounded price-assignment variation into the learned policy.

8 Discussions

We acknowledge several limitations of our work that warrant further investigation. First, the proximal causal inference framework fundamentally relies on the validity of the structural assumptions governing the treatment- and outcome-inducing confounding proxies. Violations of these assumptions may introduce bias into our proposed estimators, even in settings where unconfoundedness based on measured covariates actually holds. Therefore, the identification of valid proxies demands careful consideration, particularly in domains with limited prior knowledge. Second, although we incorporate the treatment-inducing proxy into the policy making, our identification result still requires an additional auxiliary treatment-inducing proxy, as formalized in Assumptions 4 and 5. Whether a more general class of IDRs can be estimated while preserving the computational efficiency of O-learning in continuous treatment settings remains an open question. Third, as demonstrated in our numerical studies, the performance of the estimator d_2 is sensitive to the informativeness of the proxies. A promising direction for future research involves settings with multiple treatment-inducing proxies (Yu et al., 2025). In such cases, developing an approach to select the optimal auxiliary proxy to improve model performance is of great interest.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation between Chinese and English and to improve the clarity, grammar, and phrasing of author-written text, including the wording of this disclosure. We also used these tools to identify potentially relevant literature and to summarize and analyze existing work. The authors reviewed all AI-assisted text and checked the original sources of cited works for relevance and accuracy. We did not use generative AI tools to develop the research methodology, design or conduct experiments, generate synthetic datasets, formulate mathematical claims, or assist with mathematical proofs. The authors take responsibility for the final content of this paper, including its text, claims, citations, and other artifacts.

REFERENCES

Le Thi Hoai An and Pham Dinh Tao. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of operations research*, 133(1):23–46, 2005.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Bibhas Chakraborty and Erica EM Moodie. *Statistical methods for dynamic treatment regimes*, volume 2. Springer, 2013.

Guanhua Chen, Donglin Zeng, and Michael R Kosorok. Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521, 2016.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *Journal of Business & Economic Statistics*, 44(1):67–79, 2026.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

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.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Mona Eberts, Ingo Steinwart, et al. Optimal regression rates for svms using gaussian kernels. electron. *J. Stat*, 7:1–42, 2013.

Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346(6210):1243089, 2014.

Yuanshan Gao, Yang Bai, and Yifan Cui. On multiple robustness of proximal dynamic treatment regimes. *arXiv preprint arXiv:2510.20451*, 2025.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International conference on machine learning*, pp. 1414–1423. PMLR, 2017.

Binyan Jiang, Rui Song, Jialiang Li, and Donglin Zeng. Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 29(4):1633, 2019.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pp. 1243–1251. PMLR, 2018.

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.

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6(1):263–286, 2019.

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*, pp. 7512–7523. PMLR, 2021.

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.

Erica EM Moodie, Nema Dean, and Yue Ru Sun. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243, 2014.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=RmB-88r9dL.

Zhengling Qi, Rui Miao, and Xiaoke Zhang. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, 119(546):915–928, 2024.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

James M Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.

James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pp. 69–117. Springer, 1997.

Tao Shen and Yifan Cui. Optimal treatment regimes for proximal causal learning. *Advances in Neural Information Processing Systems*, 36:47735–47748, 2023.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

Anastasios A Tsiatis, Marie Davidian, Shannon T Holloway, and Eric B Laber. *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC, 2019.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Yong Wu, Yanwei Fu, Shouyan Wang, and Xinwei Sun. Doubly robust proximal causal learning for continuous treatments. *arXiv preprint arXiv:2309.12819*, 2023.

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, 2021.

Myeonghun Yu, Xu Shi, and Eric J Tchetgen Tchetgen. Fortified proximal causal inference with many invalid proxies. *arXiv preprint arXiv:2506.13152*, 2025.

Jeffrey Zhang and Eric Tchetgen Tchetgen. On identification of optimal dynamic treatment regimes with proxies of hidden confounders. *arXiv preprint arXiv:2402.14942*, 2024.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Ying-Qi Zhao, Eric B Laber, Yang Ning, Sumona Saha, and Bruce E Sands. Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, 20(48):1–23, 2019.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

A Appendix

B Technical proofs

B.1 Supporting lemmas

Lemma 1 *(Theorem 7.23 in Steinwart & Christmann (2008)). Let $L : \mathbf{X} \times A \times \mathbb{R} \to [0, \infty)$ be a loss function. Also, let H be a separable RKHS of a measurable kernel over $\mathbf{X}$ and P be a distribution on $\mathbf{X} \times \mathbf{A}$. If the following conditions are satisfied:*

- *(A1) L satisfies the supremum bound $L(\cdot) \leq B$ for a $B > 0$.*
- *(A2) L is a locally Lipschitz continuous loss that can be clipped at $M > 0$.*
- *(A3) The variance bound $\mathbb{E}_P(L \circ \tilde{f} - L \circ f^*_{L,P})^2 \leq V \cdot (\mathbb{E}_P(L \circ \tilde{f} - L \circ f^*_{L,P}))^\nu$ is satisfied for constants $\nu \in [0, 1]$, $V \geq B^{2-\nu}$, and all $f \in H$.*
- *(A4) For fixed $n \geq 1$, there exist constants $p \in (0, 1)$ and $a \geq B$ such that the entropy number $\mathbb{E}_{D_X \sim P_X^n} e_i(id : H \to L_2(D_X)) \leq a i^{-\frac{1}{2p}}, i \geq 1$.*

Fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $L \circ f_0 \leq B_0$. Then, for all fixed $\tau > 0$ and a regularization parameter $\lambda > 0$, the SVM using H and L satisfies:

$$\lambda \|f_{D,\lambda}\|_H^2 + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \leq 9(\lambda \|f_0\|_H^2 + \mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P}) + K_0 \left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\nu+\nu p}} + 3\left(\frac{72V\tau}{n}\right)^{\frac{1}{2-\nu}} + \frac{15B_0\tau}{n},$$

with probability P^n not less than $1 - 3e^{-\tau}$.

Lemma 2 *(Theorem 7.34 in Steinwart & Christmann (2008)). Let μ be a distribution on $\mathbb{R}^d$ having tail exponent $\tau \in (0, \infty]$. Then, for all $\delta > 0$ and $\frac{d}{d+\tau} < p < 1$, there exists a constant $c_{\delta,p} \geq 1$ such that:*

$$e_i(\mathrm{id} : H_\gamma(\mathbb{R}^d) \to L_2(\mu)) \leq c_{\delta,p} \gamma^{-\frac{(1-p)(1+\delta)d}{2p}} i^{-\frac{1}{2p}},$$

for all $i \geq 1$ and $\gamma \in (0, 1]$.

Lemma 3 *(Theorem 2.2 in Eberts et al. (2013)). Let us fix some $q \in [1, \infty)$. Furthermore, assume that P_X is a distribution on $\mathbb{R}^d$ that has a Lebesgue density $g \in L_p(\mathbb{R}^d)$ for some $p \in [1, \infty]$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be such that $f \in L_q(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, and the function $K : \mathbb{R}^d \to \mathbb{R}$ be defined by:*

$$K(x) := \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2 \pi} \right)^{\frac{d}{2}} K_{\frac{j\gamma}{\sqrt{2}}}(x), \tag{8}$$

where $K_\gamma(x) := \exp(-\gamma^{-2}\|x\|_2^2)$ for all $x \in \mathbb{R}^d$. Then, for $r \in \mathbb{N}$, $\gamma > 0$, and $s \geq 1$ with $1 = \frac{1}{s} + \frac{1}{p}$, we have:

$$\|K * f - f\|_{L_q(P_X)}^q \leq C_{r,q} \|g\|_{L_p(\mathbb{R}^d)} \omega_{r, L_{qs}(\mathbb{R}^d)}^q (f, \gamma/2),$$

where $C_{r,q}$ is a constant only depending on r and q.

Lemma 4 *(Theorem 2.3 in Eberts et al. (2013)). Let $f \in L_2(\mathbb{R}^d)$, H_γ be the RKHS of the Gaussian RBF kernel k_γ over $X \subset \mathbb{R}^d$ with $\gamma > 0$ and $K : \mathbb{R}^d \to \mathbb{R}$ be defined by Equation 8 for a fixed $r \in \mathbb{N}$. Then we have $K * f \in H_\gamma$ with:*

$$\|K * f\|_{H_\gamma} \leq (\gamma \sqrt{\pi})^{-\frac{d}{2}} (2^r - 1) \|f\|_{L_2(\mathbb{R}^d)}.$$

B.2 Proofs for section 3

Proof of theorem 1: Note that:

$$\begin{aligned}
&\lim_{\phi \to 0^+} \mathbb{E}\left[\frac{Y \mathbb{I}(|A - d_1(X,W)| \leq \phi)}{2\phi} q_0(Z, A, X) \,\middle|\, U, X, W \right] \\
&= \lim_{\phi \to 0^+} \mathbb{E}\left[\frac{\mathbb{I}(|A - d_1(X,W)| \leq \phi)}{2\phi} \mathbb{E}[Y|A, U, X, W] \frac{1}{p(A|U,X)} \,\middle|\, U, X, W \right] \\
&= \lim_{\phi \to 0^+} \int_{d_1(X,W)-\phi}^{d_1(X,W)+\phi} \frac{1}{2\phi} \frac{p(a|U,X)}{p(a|U,X)} \mathbb{E}[Y|A = a, U, X, W] \, \mathrm{d}a \\
&= \mathbb{E}[Y(d_1(X,W))|U, X, W].
\end{aligned}$$

where the first equality is implied by Theorem 2.2 of Cui et al. (2024), and the last equality follows from the continuity of $\mathbb{E}[Y(a)|U, X, W]$ with respect to a.

Since Y is assumed to be bounded, the desired result is obtained by applying the dominated convergence theorem to the expectation over (U, X, W).

Proof of Theorem 2: Figure 3 depicts the graphical structure used in this proof.

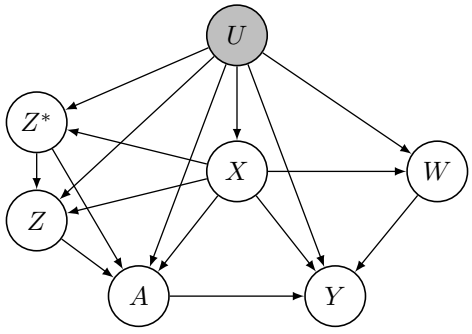


Figure 3: DAG for Theorem 2 with auxiliary treatment-inducing proxy Z^*.

We first show that

$$\mathbb{E}\{q_1(Z^*, A, X) \mid Z, X, U, A\} = \frac{1}{p(A \mid Z, X, U)}.$$

To this end, note that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{p(A \mid U, X, Z)} \mid W, A, X, Z\right] \\
=&\int \frac{1}{p(A \mid u, X, Z)}\, dF(u \mid W, A, X, Z) \\
=&\int \frac{1}{p(A \mid u, X, Z, W)}\, dF(u \mid W, A, X, Z) \\
=&\mathbb{E}\left[q_1(Z^*, A, X) \mid W, A, X, Z\right] \\
=&\mathbb{E}\left[\mathbb{E}\{q_1(Z^*, A, X) \mid U, W, A, X, Z\} \mid W, A, X, Z\right] \\
=&\mathbb{E}\left[\mathbb{E}\{q_1(Z^*, A, X) \mid U, A, X, Z\} \mid W, A, X, Z\right].
\end{aligned}
$$

The third equality follows from Assumption 5. For the last equality, by Assumption 4 we have $Z^* \perp\!\!\!\perp W \mid (U, X, A, Z)$, hence $\mathbb{E}[q_1(Z^*, A, X) \mid U, W, A, X, Z] = \mathbb{E}[q_1(Z^*, A, X) \mid U, A, X, Z]$. Therefore, by the completeness condition in Assumption 6,

$$
\mathbb{E}\{q_1(Z^*, A, X) \mid U, A, X, Z\} = \frac{1}{p(A \mid U, X, Z)}.
$$

We now establish the desired identification result. For any rule d_2, we have

$$
\begin{aligned}
&\lim_{\phi \to 0^+} \mathbb{E}\left[\frac{Y \mathbb{I}(|A - d_2(X, W, Z)| \le \phi)}{2\phi} q_1(Z^*, A, X) \Big| U, X, W, Z\right] \\
=&\lim_{\phi \to 0^+} \mathbb{E}\left[\frac{\mathbb{E}[Y|A, U, X, W, Z]\mathbb{I}(|A - d_2(X, W, Z)| \le \phi)}{2\phi} \mathbb{E}[q_1(Z^*, A, X)|A, U, X, W, Z] \Big| U, X, W, Z\right] \\
=&\lim_{\phi \to 0^+} \mathbb{E}\left[\frac{\mathbb{E}[Y|A, U, X, W, Z]\mathbb{I}(|A - d_2(X, W, Z)| \le \phi)}{2\phi} \frac{1}{p(A|U, X, W, Z)} \Big| U, X, W, Z\right] \\
=&\lim_{\phi \to 0^+} \int_{d_2(X,W,Z)-\phi}^{d_2(X,W,Z)+\phi} \frac{1}{2\phi} \mathbb{E}[Y \Big| A = a, U, X, W, Z] \frac{p(a|U, X, W, Z)}{p(a|U, X, W, Z)} \mathrm{d}a \\
=&\lim_{\phi \to 0^+} \int_{d_2(X,W,Z)-\phi}^{d_2(X,W,Z)+\phi} \frac{1}{2\phi} \mathbb{E}[Y(a)|U, X, W, Z] \mathrm{d}a \\
=&\mathbb{E}[Y(d_2(X, W, Z))|U, X, W, Z],
\end{aligned}
$$

where the first equality follows from $Z^* \perp\!\!\!\perp Y \mid (A, U, X, W, Z)$, implied by Assumption 4. The second equality uses $Z^* \perp\!\!\!\perp W \mid (U, X, A, Z)$ from the same Assumption 4, the established identity $\mathbb{E}[q_1(Z^*, A, X) \mid Z, X, U, A] = 1/p(A \mid Z, X, U)$, and $A \perp\!\!\!\perp W \mid (U, X, Z)$. The fourth equality follows from consistency and $A \perp\!\!\!\perp Y(a) \mid (U, X, W, Z)$, which follows by weak union from the joint independence $(Z, A) \perp\!\!\!\perp (Y(a), W) \mid (U, X)$ in Assumption 1(3). The final equality follows from the continuity of $\mathbb{E}[Y(a) \mid U, X, W, Z]$ with respect to a.

Due to the fact that Y is bounded, we obtain the desired result by the dominated convergence theorem.

B.3 PROOFS FOR SECTION 5

Proof of theorem 3: First, we have:

$$
\begin{aligned}
V_{\phi_n}(d_2) &= E\left\{\frac{Yq_1(Z^*,A,X)}{\phi_n}\max(1-\frac{|A-d_2(X,W,Z)|}{\phi_n},0)\right\} \\
&= E_{(U,X,W,Z)}\left\{\frac{1}{\phi_n^2}\int_{|d_2(X,W,Z)-a|\le\phi_n}[\phi_n-|a-d_2(X,W,Z)|]E(Y\mid A=a,U,X,W,Z)\mathrm{d}a\right\} \\
&= E_{(U,X,W,Z)}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]E(Y\mid A=z+d_2(X,W,Z),U,X,W,Z)\mathrm{d}z\right\} \\
&= E_{(U,X,W,Z)}\Big\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]\big[E(Y\mid A=d_2(X,W,Z),U,X,W,Z) \\
&+ z\frac{\partial E(Y\mid A=a,U,X,W,Z)}{\partial a}\mid_{a=d_2(X,W,Z)}+o(z)\big]\mathrm{d}z\Big\} \\
&= V(d_2)+E_{(U,X,W,Z)}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]\left[z\frac{\partial E(Y\mid A=a,U,X,W,Z)}{\partial a}\mid_{a=d_2(X,W,Z)}+o(z)\right]\mathrm{d}z\right\}.
\end{aligned}
$$

Therefore,

$$
|V_{\phi_n}(d_2)-V(d_2)| = \left|E_{(U,X,W,Z)}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]\left[z\frac{\partial E(Y\mid A=a,U,X,W,Z)}{\partial a}\mid_{a=d_2(X,W,Z)}+o(z)\right]\mathrm{d}z\right\}\right|.
$$

By the Lipschitz condition, we conclude that

$$
|V_{\phi_n}(d_2)-V(d_2)| \le C\int_{|z|\le\phi_n}\frac{\phi_n|z|-z^2}{\phi_n^2}dz = C\phi_n.
$$

Proof of theorem 4

Let $d^*_{\phi_n}$ be the minimizer of $\mathscr{R}_{\phi_n}(d_2,\hat{q}_1)$, and by definition $\mathscr{R}_{\phi_n}(d_2,\hat{q}_1) = \mathbb{E}\left(\frac{Y\hat{q}_1(Z^*,A,X)}{\phi_n}\ell_{\phi_n}(A-d_2(X,W,Z))\right)$. Recall that the surrogate value function is defined as $V_{\phi_n}(d_2)=\mathbb{E}\left[\frac{Yq_1(Z^*,A,X)}{\phi_n}\right]-\mathscr{R}_{\phi_n}(d_2,q_1)$. Then

$$
\begin{aligned}
V(d_2^*)-V(\hat{d}_2) &\le V_{\phi_n}(d_2^*,q_1)-V_{\phi_n}(\hat{d}_2,q_1)+2C\phi_n \\
&\le V_{\phi_n}(d_2^*,\hat{q}_1)-V_{\phi_n}(\hat{d}_2,\hat{q}_1)+|V_{\phi_n}(\hat{d}_2,q_1)-V_{\phi_n}(\hat{d}_2,\hat{q}_1)| \\
&+|V_{\phi_n}(d_2^*,q_1)-V_{\phi_n}(d_2^*,\hat{q}_1)|+2C\phi_n \\
&= \mathscr{R}_{\phi_n}(\hat{d}_2,\hat{q}_1)-\mathscr{R}_{\phi_n}(d_2^*,\hat{q}_1)+|V_{\phi_n}(\hat{d}_2,q_1)-V_{\phi_n}(\hat{d}_2,\hat{q}_1)| \\
&+|V_{\phi_n}(d_2^*,q_1)-V_{\phi_n}(d_2^*,\hat{q}_1)|+2C\phi_n \\
&\le (\mathscr{R}_{\phi_n}(\hat{d}_2,\hat{q}_1)-\mathscr{R}_{\phi_n}(d^*_{\phi_n},\hat{q}_1))+|V_{\phi_n}(\hat{d}_2,q_1)-V_{\phi_n}(\hat{d}_2,\hat{q}_1)| \\
&+|V_{\phi_n}(d_2^*,q_1)-V_{\phi_n}(d_2^*,\hat{q}_1)|+c_7\phi_n.
\end{aligned}
\tag{9}
$$

The second inequality is due to Theorem 3, and the third follows by the definition of $\mathscr{R}_{\phi_n}$. The final inequality holds by the definition of $d^*_{\phi_n}$, where C is the same constant used in Theorem 3 and $c_7=2C$.

To bound the first term in Equation 9, we proceed in two steps.

Step I. Let $\mathscr{R}^*_{\phi_n} := \mathscr{R}_{\phi_n}(d^*_{\phi_n},\hat{q}_1)$ denote this minimum risk. To bound the excess risk $\mathscr{R}_{\phi_n}(\hat{d}_2,\hat{q}_1)-\mathscr{R}^*_{\phi_n}$, define the signed scaled criterion $L_{\phi_n}(O,d_2) := Y\hat{q}_1(Z^*,A,X)\ell_{\phi_n}(A-d_2(X,W,Z))$, where $O=(Y,A,X,W,Z,Z^*)$, and let $\mathcal{R}_{L,P}(d_2) := \mathbb{E}[L_{\phi_n}(\cdot,d_2)] = \phi_n\mathscr{R}_{\phi_n}(d_2,\hat{q}_1)$. Let $G_{\hat{q}}(O)=Y\hat{q}_1(Z^*,A,X)$, and define:

$$
\bar{L}_{\phi_n}(O,d_2)=G^+_{\hat{q}}(O)\ell_{\phi_n}(A-d_2(X,W,Z))+G^-_{\hat{q}}(O)\{1-\ell_{\phi_n}(A-d_2(X,W,Z))\}.
$$

Then

$$\bar{L}_{\phi_n}(O, d_2) = L_{\phi_n}(O, d_2) + G^-_{\hat{q}}(O),$$

where the second term is independent of d_2. Hence the empirical and population risks induced by $\bar{L}_{\phi_n}$ differ from those induced by the original signed criterion only by constants independent of d_2. The regularized empirical minimizer and all excess risk differences are therefore unchanged, and $d^*_{\phi_n}$ is also the minimizer of $\bar{\mathcal{R}}_{L,P}$, Let $\bar{\mathcal{R}}_{L,P}(d_2) := \mathbb{E}[\bar{L}_{\phi_n}(\cdot, d_2)]$. We have:

$$\mathscr{R}_{\phi_n}(\hat{d}_2, \hat{q}_1) - \mathscr{R}^*_{\phi_n} \leq \frac{1}{\phi_n}\left[\lambda_n \phi_n \|\hat{d}_2\|^2_{\mathcal{D}_2} + \bar{\mathcal{R}}_{L,P}(\hat{d}_2) - \bar{\mathcal{R}}_{L,P}(d^*_{\phi_n})\right]. \tag{10}$$

We pre-estimate the nuisance function estimator $\hat{q}_1$ via sample splitting independent of the policy learning, we now verify the conditions of Lemma 1 for the nonnegative scaled loss $\bar{L}_{\phi_n}$ and risk $\bar{\mathcal{R}}_{L,P}$.

For (A1), the boundedness in Assumptions 7 and 8 imply $|G_{\hat{q}}(O)| \leq B$ for some constant $B > 0$. Since $0 \leq \ell_{\phi_n} \leq 1$, the loss satisfies $0 \leq \bar{L}_{\phi_n}(\cdot) \leq B$.

For (A2), since $\mathcal{A}$ is bounded, we choose $M \geq \sup_{a\in\mathcal{A}} |a|$ so that clipping acts as the identity on the policy class, and thus the clipping condition holds. Moreover, $\bar{L}_{\phi_n}(M^{(2)}, A, d_2(X, W, Z))$ is Lipschitz continuous with respect to d_2 with Lipschitz constant B/ϕ_n, and is therefore also locally Lipschitz continuous.

For (A3), we can bound the variance for any $d_2 \in \mathcal{D}_2$ as follows:

$$\mathbb{E}_P(\bar{L}_{\phi_n} \circ \tilde{d}_2 - \bar{L}_{\phi_n} \circ d^*_{\phi_n})^2 \leq 2\mathbb{E}_P\left[\left(\bar{L}_{\phi_n} \circ \tilde{d}_2\right)^2 + \left(\bar{L}_{\phi_n} \circ d^*_{\phi_n}\right)^2\right] \leq 4B^2.$$

The last inequality follows from (A1). Therefore, the condition is satisfied with parameters $\nu = 0$ and $V = 4B^2$.

For (A4), since we use the Gaussian kernel as the benign kernel, and distribution $P_{(X,W,Z)}$ has tail exponent $t \in (0, \infty]$, Lemma 2 guarantees that this condition holds, with the constant a scaling as $a = c_{\delta,p}\gamma_n^{-\frac{(1-p)(1+\delta)d}{2p}}$.

With all conditions satisfied, applying Lemma 1 yields that with probability at least $1 - 3e^{-\tau}$:

$$\begin{aligned}\lambda_n \phi_n \|\hat{d}_2\|^2_{\mathcal{D}_2} + \bar{\mathcal{R}}_{L,P}(\hat{d}_2) - \bar{\mathcal{R}}_{L,P}(d^*_{\phi_n}) \leq 9A(\lambda_n) + K_0\left[\frac{1}{\gamma_n^{(1-p)(1+\delta)d}\lambda_n^p \phi_n^p n}\right]^{\frac{1}{2-p}} \\ + 36\sqrt{2}B\left(\frac{\tau}{n}\right)^{\frac{1}{2}} + \frac{15B\tau}{n},\end{aligned} \tag{11}$$

where $A(\lambda_n) := \lambda_n \phi_n \|d_2\|^2_{\mathcal{D}_2} + \bar{\mathcal{R}}_{L,P}(d_2) - \bar{\mathcal{R}}_{L,P}(d^*_{\phi_n})$.

Step II. To bound the approximation error $A(\lambda_n)$, recall $\mathcal{D}_2$ is taken as the RKHS associated with Gaussian kernel of bandwidth γ_n, since any $d_2 \in \mathcal{D}_2$ is valid for Equation 11, we can study a specific choice of d_2 and the corresponding bound for $A(\lambda_n)$. To construct this d_2, for $r \in \mathbb{N}$ and $\gamma > 0$, we define the function $K : \mathbb{R}^d \to \mathbb{R}$ as:

$$K(x) = \sum_{j=1}^{r} \binom{r}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2 \pi}\right)^{d/2} K_{\frac{j\gamma}{\sqrt{2}}}(x),$$

where $K_\gamma = \exp(-\gamma^{-2}\|x\|_2^2)$ for all $x \in \mathbb{R}^d$. Assuming that $d^*_{\phi_n} \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, we define d_2 by convolving K with this optimal decision function, that is:

$$d_2 := K * d^*_{\phi_n} = \int_{\mathbb{R}^d} K(x - t) d^*_{\phi_n}(t) dt, \quad x \in \mathbb{R}^d.$$

By Lemma 4, this construction ensures $d_2 \in \mathcal{D}_2$, and we have:

$$\begin{aligned}
A(\lambda_n) &= \lambda_n \phi_n \|d_2\|_{\mathcal{D}_2}^2 + \bar{\mathcal{R}}_{L,P}(d_2) - \bar{\mathcal{R}}_{L,P}(d_{\phi_n}^*) = \lambda_n \phi_n \|K * d_{\phi_n}^*\|_{\mathcal{D}_2}^2 + \bar{\mathcal{R}}_{L,P}(K * d_{\phi_n}^*) - \bar{\mathcal{R}}_{L,P}(d_{\phi_n}^*) \\
&\leq \lambda_n \phi_n (\gamma_n \sqrt{\pi})^{-d} (2^r - 1)^2 \|d_{\phi_n}^*\|_{L_2(\mathbb{R}^d)}^2 + \bar{\mathcal{R}}_{L,P}(K * d_{\phi_n}^*) - \bar{\mathcal{R}}_{L,P}(d_{\phi_n}^*) \quad \text{(Lemma 4)} \\
&\leq \lambda_n \phi_n (\gamma_n \sqrt{\pi})^{-d} (2^r - 1)^2 \|d_{\phi_n}^*\|_{L_2(\mathbb{R}^d)}^2 + \frac{B}{\phi_n} \|K * d_{\phi_n}^* - d_{\phi_n}^*\|_{L_1(P_X)} \quad \text{(Lipschitz property of } \bar{L}_{\phi_n}) \\
&\leq \lambda_n \phi_n (\gamma_n \sqrt{\pi})^{-d} (2^r - 1)^2 \|d_{\phi_n}^*\|_{L_2(\mathbb{R}^d)}^2 + \frac{B}{\phi_n} C_{r,1} \|g\|_{L_p(\mathbb{R}^d)} \omega_{r,L_1(\mathbb{R}^d)}(d_{\phi_n}^*, \frac{\gamma_n}{2}). \quad \text{(Lemma 3)}
\end{aligned} \tag{12}$$

Recall that we assume $d_2^* \in B_{1,\infty}^{\alpha}(\mathbb{R}^d)$, and if we further assume $d_{\phi_n}^* \in B_{1,\infty}^{\alpha}(\mathbb{R}^d)$, i.e., $B_{1,\infty}^{\alpha}(\mathbb{R}^d) = \{f \in L_\infty(\mathbb{R}^d) : \sup_{t>0}(t^{-\alpha}\omega_{r,L_1(\mathbb{R}^d)}(f,t)) < \infty\}$, then $\omega_{r,L_1(\mathbb{R}^d)}(d_{\phi_n}^*, \frac{\gamma_n}{2}) < c_0 \gamma_n^{\alpha}$ where c_0 is a constant. Plugging this into inequality Equation 12, and combining with the assumption $d_{\phi_n}^* \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, we obtain:

$$A(\lambda_n) \leq c_1 \lambda_n \phi_n \gamma_n^{-d} + c_2 \frac{\gamma_n^{\alpha}}{\phi_n}.$$

Substituting this bound for $A(\lambda_n)$ into Equation 11 and then plugging the result back into Equation 10, we obtain:

$$\mathscr{R}_{\phi_n}(\hat{d}_2, \hat{q}_1) - \mathscr{R}_{\phi_n}^* \leq c_1 \left[\frac{1}{\gamma_n^{(1-p)(1+\delta)d} \lambda_n^p \phi_n^2 n} \right]^{\frac{1}{2-p}} + c_2 \frac{\tau^{1/2}}{\phi_n n^{1/2}} + c_3 \frac{\tau}{\phi_n n} + c_4 \frac{\lambda_n}{\gamma_n^d} + c_5 \frac{\gamma_n^{\alpha}}{\phi_n^2}.$$

To bound the second term in Equation 9, we apply the Cauchy-Schwarz inequality. Specifically, we evaluate the error introduced by the estimation of the nuisance function q_1:

$$\begin{aligned}
&|V_{\phi_n}(\hat{d}_2, q_1) - V_{\phi_n}(\hat{d}_2, \hat{q}_1)| \\
\leq &\mathbb{E}\left[\frac{|Y|}{\phi_n} \max\left(1 - \frac{|A - d_2(X,W,Z)|}{\phi_n}, 0\right) |q_1(Z^*, A, X) - \hat{q}_1(Z^*, A, X)| \right] \\
\leq &\frac{C_1}{\phi_n} \sqrt{\mathbb{E}\left[\max\left(1 - \frac{|A - d_2(X,W,Z)|}{\phi_n}, 0\right)^2 \right]} \cdot \|\hat{q}_1 - q_1\|_2 \\
\leq &\frac{C_1}{\phi_n} \sqrt{\mathbb{E}_{(X,W,Z)}\left[\mathbb{E}\left[\int_{d_2(X,W,Z)-\phi_n}^{d_2(X,W,Z)+\phi_n} p(a|X,W,Z) \mathrm{d}a \right] \right]} \cdot \|\hat{q}_1 - q_1\|_2 \\
= &O_p\left(\frac{1}{\sqrt{\phi_n}} \right) \cdot \|\hat{q}_1 - q_1\|_2 \\
= &O_p\left(\frac{n^{-\varphi}}{\sqrt{\phi_n}} \right).
\end{aligned}$$

Combining the above yields the final result, which holds with probability at least $1 - 3e^{-\tau}$:

$$\begin{aligned}
V(d_2^*) - V(\hat{d}_2) \leq c_1 &\left[\frac{1}{\gamma_n^{(1-p)(1+\delta)d} \lambda_n^p \phi_n^2 n} \right]^{\frac{1}{2-p}} + c_2 \frac{\tau^{\frac{1}{2}}}{\phi_n n^{\frac{1}{2}}} + c_3 \frac{\tau}{\phi_n n} \\
&+ c_4 \frac{\lambda_n}{\gamma_n^d} + c_5 \frac{\gamma_n^{\alpha}}{\phi_n^2} + c_6 \frac{n^{-\varphi}}{\sqrt{\phi_n}} + c_7 \phi_n.
\end{aligned}$$

C COMPUTATIONAL DETAILS OF THE DC ALGORITHM

In this section, we provide a detailed derivation to demonstrate that the optimization subproblem in the DC algorithm is equivalent to a quadratic program. Recall that $w_i = Y_i \hat{q}_0(Z_i, A_i, X_i)$, $\mathcal{P} = \{i : w_i > 0\}$ and $\mathcal{N} = \{i : w_i < 0\}$, thus $w_i^+ = w_i$ for $i \in \mathcal{P}$ and $w_i^- = -w_i$ for $i \in \mathcal{N}$. Then

$S(\Theta) = S_1(\Theta) - S_2(\Theta)$, where:

$$S_1(\Theta) = \frac{\lambda_n}{2}\|\beta\|_2^2 + \frac{1}{n\phi_n^2}\sum_{i\in\mathcal{P}} w_i^+ |A_i - (\beta^\top M_i^{(1)} + b)| + \frac{1}{n\phi_n}\sum_{i\in\mathcal{N}} w_i^- \left(\frac{|A_i - (\beta^\top M_i^{(1)} + b)|}{\phi_n} - 1\right)_+,$$

$$S_2(\Theta) = \frac{1}{n\phi_n^2}\sum_{i\in\mathcal{N}} w_i^- |A_i - (\beta^\top M_i^{(1)} + b)| + \frac{1}{n\phi_n}\sum_{i\in\mathcal{P}} w_i^+ \left(\frac{|A_i - (\beta^\top M_i^{(1)} + b)|}{\phi_n} - 1\right)_+.$$

The DC algorithm proceeds by initializing Θ_0 and iteratively solving a convex majorization of the objective. At each iteration t, we linearize the concave part $-S_2(\Theta)$ at the current estimate Θ_t and solve the following convex subproblem:

$$\Theta_{t+1} = \arg\min_{\Theta} \left(S_1(\Theta) - \left[S_2(\Theta_t) + \nabla S_2(\Theta_t)^T(\Theta - \Theta_t)\right]\right). \tag{13}$$

At iteration t, the objective function in Equation 13 can be expressed as:

$$\begin{aligned}\min_{\beta,b} \quad & \frac{\lambda_n}{2}\|\beta\|_2^2 + \frac{1}{n\phi_n^2}\sum_{i\in\mathcal{P}} w_i^+ |A_i - (\beta^\top M_i^{(1)} + b)| + \frac{1}{n\phi_n}\sum_{i\in\mathcal{N}} w_i^- \left(\frac{|A_i - (\beta^\top M_i^{(1)} + b)|}{\phi_n} - 1\right)_+ \\ & - \beta^\top \nabla_\beta S_2(\Theta_t) - b\nabla_b S_2(\Theta_t).\end{aligned} \tag{14}$$

Dividing Equation 14 by λ_n, denote $s_i^t = \text{sign}(A_i - (\beta_t^\top M_i^{(1)} + b_t))$, define

$$C_i = \frac{|w_i|}{n\lambda_n\phi_n^2}, \qquad h_i^{(t)} = C_i s_i^t \mathbb{I}\{i \in \mathcal{N} \text{ or } (i \in \mathcal{P},\ Q_i^{(t)} = 0)\}.$$

After omitting constants independent of (β, b), the exact convex majorization is

$$\begin{aligned}\min_{\beta,b} \quad & \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^n h_i^{(t)}(\beta^\top M_i^{(1)} + b) + \sum_{i\in\mathcal{P}} C_i |A_i - (\beta^\top M_i^{(1)} + b)| \\ & + \sum_{i\in\mathcal{N}} C_i \big(|A_i - (\beta^\top M_i^{(1)} + b)| - \phi_n\big)_+.\end{aligned} \tag{15}$$

Let $f_i = \beta^\top M_i^{(1)} + b$. For $i \in \mathcal{P}$, split the residual into its positive and negative parts,

$$\xi_i = (A_i - f_i)_+, \qquad \xi_i^* = (A_i - f_i)_-.$$

Thus, $A_i - f_i = \xi_i - \xi_i^*$ and $|A_i - f_i| = \xi_i + \xi_i^*$. When ξ_i and ξ_i^* are treated as optimization variables, these identities are enforced by $A_i - f_i = \xi_i - \xi_i^*$ and $\xi_i, \xi_i^* \geq 0$: because $C_i > 0$, every optimum has at most one of ξ_i and ξ_i^* strictly positive. For $i \in \mathcal{N}$, introduce the two ϕ_n-insensitive slack variables

$$\zeta_i = (A_i - f_i - \phi_n)_+, \qquad \zeta_i^* = (-(A_i - f_i) - \phi_n)_+,$$

so that $(|A_i - f_i| - \phi_n)_+ = \zeta_i + \zeta_i^*$ at the optimum. The primal QP for the $(t+1)$-th iteration can therefore be written as

$$\begin{aligned}\min_{\beta,b,\xi,\xi^*,\zeta,\zeta^*} \quad & \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^n h_i^{(t)} f_i + \sum_{i\in\mathcal{P}} C_i(\xi_i + \xi_i^*) + \sum_{i\in\mathcal{N}} C_i(\zeta_i + \zeta_i^*) \\ \text{s.t.} \quad & A_i - f_i = \xi_i - \xi_i^*, \quad \xi_i \geq 0, \quad \xi_i^* \geq 0, \quad i \in \mathcal{P}, \\ & A_i - f_i - \phi_n \leq \zeta_i, \quad f_i - A_i - \phi_n \leq \zeta_i^*, \quad i \in \mathcal{N}, \\ & \zeta_i \geq 0, \quad \zeta_i^* \geq 0, \quad i \in \mathcal{N}.\end{aligned}$$

To derive the dual, associate an unrestricted multiplier α_i with the equality constraint for $i \in \mathcal{P}$, multipliers $\mu_i, \mu_i^* \geq 0$ with the two insensitive-loss constraints for $i \in \mathcal{N}$, and multipliers $\kappa_i, \kappa_i^*, \rho_i, \rho_i^* \geq 0$ with the nonnegativity constraints on $\xi_i, \xi_i^*, \zeta_i, \zeta_i^*$, respectively. The Lagrangian

is

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^{n} h_i^{(t)} f_i + \sum_{i\in\mathcal{P}} C_i(\xi_i+\xi_i^*) + \sum_{i\in\mathcal{N}} C_i(\zeta_i+\zeta_i^*) \\ + \sum_{i\in\mathcal{P}} \alpha_i[A_i - f_i - \xi_i + \xi_i^*] \\ - \sum_{i\in\mathcal{N}} \mu_i[\zeta_i - A_i + f_i + \phi_n] - \sum_{i\in\mathcal{N}} \mu_i^*[\zeta_i^* + A_i - f_i + \phi_n] \\ - \sum_{i\in\mathcal{P}} (\kappa_i\xi_i + \kappa_i^*\xi_i^*) - \sum_{i\in\mathcal{N}} (\rho_i\zeta_i + \rho_i^*\zeta_i^*).\end{aligned}$$

Define the unified dual variables

$$\eta_i = \begin{cases}\alpha_i, & i\in\mathcal{P},\\ \mu_i - \mu_i^*, & i\in\mathcal{N},\end{cases} \qquad v_i = \mu_i + \mu_i^*, \quad i\in\mathcal{N}.$$

The KKT stationarity conditions give

$$\frac{\partial\mathcal{L}}{\partial\beta} = \beta + \sum_{i=1}^{n}(h_i^{(t)} - \eta_i)M_i^{(1)} = 0 \quad\Longrightarrow\quad \beta = \sum_{i=1}^{n}(\eta_i - h_i^{(t)})M_i^{(1)},$$

$$\frac{\partial\mathcal{L}}{\partial b} = \sum_{i=1}^{n}(h_i^{(t)} - \eta_i) = 0,$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i} = C_i - \alpha_i - \kappa_i = 0, \qquad \frac{\partial\mathcal{L}}{\partial\xi_i^*} = C_i + \alpha_i - \kappa_i^* = 0, \quad i\in\mathcal{P},$$

$$\frac{\partial\mathcal{L}}{\partial\zeta_i} = C_i - \mu_i - \rho_i = 0, \qquad \frac{\partial\mathcal{L}}{\partial\zeta_i^*} = C_i - \mu_i^* - \rho_i^* = 0, \quad i\in\mathcal{N}.$$

Hence, $-C_i \le \eta_i \le C_i$ for $i\in\mathcal{P}$. While $0 \le \mu_i, \mu_i^* \le C_i$ for $i\in\mathcal{N}$. Under the change of variables $(\mu_i, \mu_i^*) \mapsto (\eta_i, v_i)$ equivalently we have

$$v_i - \eta_i \ge 0, \quad v_i + \eta_i \ge 0, \quad v_i - \eta_i \le 2C_i, \quad v_i + \eta_i \le 2C_i, \qquad i\in\mathcal{N}.$$

In particular, $-C_i \le \eta_i \le C_i$ also holds for $i\in\mathcal{N}$. Denote $K_{ij} = \langle M_i^{(1)}, M_j^{(1)}\rangle$, substituting β into the Lagrangian yields

$$\begin{aligned}\min_{\eta,v} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\eta_i\eta_j K_{ij} - \sum_{i=1}^{n}\eta_i\left(A_i + \sum_{j=1}^{n} h_j^{(t)} K_{ij}\right) + \phi_n\sum_{i\in\mathcal{N}} v_i \\ \text{s.t.} \quad & \sum_{i=1}^{n}\eta_i = \sum_{i=1}^{n} h_i^{(t)}, \\ & -C_i \le \eta_i \le C_i, \quad i = 1,\dots,n, \\ & v_i - \eta_i \ge 0, \quad v_i + \eta_i \ge 0, \quad i\in\mathcal{N}.\end{aligned}$$

The two omitted upper constraints $v_i - \eta_i \le 2C_i$ and $v_i + \eta_i \le 2C_i$ do not change the optimum of this simplified QP. Indeed, because $\phi_n > 0$, minimizing over v_i for fixed η_i gives $v_i = |\eta_i|$; together with $|\eta_i| \le C_i$, this value satisfies the two upper constraints. The primal slope is recovered from $\beta_{t+1} = \sum_i(\eta_i - h_i^{(t)})M_i^{(1)}$. For this fixed slope, the intercept is recovered exactly from the one-dimensional convex problem

$$\min_b \sum_{i=1}^{n} h_i^{(t)}(M_i^{(1)\top}\beta_{t+1} + b) + \sum_{i\in\mathcal{P}} C_i|A_i - M_i^{(1)\top}\beta_{t+1} - b| + \sum_{i\in\mathcal{N}} C_i\big(|A_i - M_i^{(1)\top}\beta_{t+1} - b| - \phi_n\big)_+.$$

D Additional experiment details

D.1 Data-generating mechanisms

We generate $U_2 \sim \text{Uniform}[-1, 2]$ and $U_1 \sim \text{Uniform}[0, 1] - \mathbb{I}(0 \le U_2 \le 1)$. The proxies are constructed as $W = [U_1 + \text{Uniform}[-1, 1], U_2 + \mathcal{N}(0, \sigma^2)]^\top$ and $Z = [U_1 + \mathcal{N}(0, \sigma^2), U_2 +$

Uniform$[-1,1]]^\top$. The covariates are $X = [X_1, X_2, X_3]^\top$, where $X_j \overset{\text{i.i.d.}}{\sim}$ Uniform$[-1,1]$. The treatment is $A = U_2 + \mathcal{N}(0, \sigma^2)$, where we set $\sigma^2 = 1$.

Scenario 1: The outcome Y is generated by adding a random noise from $\mathcal{N}(0, 2^2)$ to:

$$\mathbb{E}\left[Y \mid U, A, X, W, Z\right] = 4X_1 - 2X_2 - 2X_3 + 7U_1 + 10\exp\left(-\frac{(A - d^*(X,W))^4}{0.1}\right), \quad (16)$$

with oracle optimal rule:

$$d^*(X, W) = 1 + 0.5W_1 + 0.5W_2. \quad (17)$$

Scenario 2: The outcome Y is generated by adding a random noise from $\mathcal{N}(0, 2^2)$ to:

$$\mathbb{E}\left[Y \mid U, A, X, W, Z\right] = 2\cos(2\pi X_2) - 4X_3^3 + 4U_2 + 10\exp\left(-\frac{(A - d^*(X,W))^4}{0.1}\right), \quad (18)$$

with oracle optimal rule:

$$d^*(X, W) = 0.5W_1^2 + 0.5\log(|W_2| + 1) - 1. \quad (19)$$

D.2 Hyperparameters selection

In all our studies, the RKHSs $\mathcal{D}_1, \mathcal{D}_2, \mathcal{Q}, \mathcal{M}$ are equipped with Gaussian kernels of the following form:

$$K(x_1, x_2) = \exp\{-\gamma\|x_1 - x_2\|_2^2\}.$$

The bandwidth parameter is determined using the median heuristic, where $\gamma^{-1} = \text{median}\{\|x_i - x_j\|_2^2\}_{i<j\in I}$ for indices subset $I \subset \{1, \ldots, n\}$.

For CNFs, we employ the RealNVP flow structure with 2 stacked coupling layers to estimate the density.

Table 2: CNFs-block

Layer	Configuration
1	Input(A_i, W_i, Z_i, X_i)
2	FC(in-dim, 64), Tanh
3	FC(64, 32)
4	Stack 2 RealNVP coupling layers

Table 3: Hyperparameters for CNFs

Hyperparameter	
Learning rate	1e-3
Epochs	500
Batch size	128
Weight decay	1e-4

For the simulation baselines, PMMR uses product Gaussian kernels with coordinatewise median bandwidths, where the conditioning-side bandwidth is divided by 0.6, the kernel amplitude is set to 25^2, and the Nyström rank is $\min(300, n)$. CPO uses a boundary-normalized Epanechnikov kernel with bandwidth $2.6(300/n)^{1/5}$ and CNF inverse-density weights capped at 10, a ridge penalty of 10^{-4}, and 3 L-BFGS-B starts. CPO-DR uses regression trees, whereas CPO-IPW uses self-normalized weights. VCNet-TR uses quadratic dose-varying layers and is trained with SGD; its numerical settings are summarized in Table 4. Both PMMR and VCNet-TR use a 101-point dose grid.

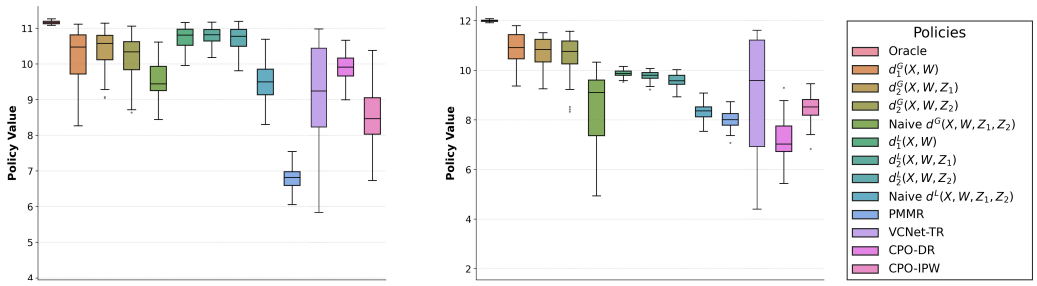
Table 4: Numerical hyperparameters for VCNet-TR

Hyperparameter	Value
Representation layers	2×50 units
Dose-varying layer knots	$(0.33, 0.66)$
Epochs	500
Batch size	500
Learning rate (network)	10^{-4}
Learning rate (targeted correction)	10^{-3}
Weight decay	5×10^{-3}
Loss weights	$\alpha = 0.5,\ \beta = 1$

In the DC algorithm, the bandwidth parameter ϕ_n is set as $\phi_n = c\hat{\sigma}_A n^{-1/6}$, where $\hat{\sigma}_A$ is the estimated standard deviation of A. The hyperparameter $c > 0$ is selected from the candidate set $\mathcal{C} = \{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$. Specifically, we employ a 3-fold partitioning for training and validation. For each $c \in \mathcal{C}$, the regularization parameters λ_n and γ_n are optimized via 5-fold cross-validation on the training set. With a fixed small $\varepsilon = 0.1$, we estimate the optimal IDR $\hat{d}^{(-k)}$ on the training set and evaluate the empirical value $V_\varepsilon(\hat{d}^{(-k)})$ on the validation set. The optimal c is chosen to maximize the average predicted value across all validation folds. The final tuning parameters λ_n and γ_n are then selected through another 5-fold cross-validation on the entire dataset.

To accelerate computation, Gaussian kernels for our proposed estimators are approximated using $2\lceil\sqrt{n}\rceil$ basis functions via the Nyström method.

D.3 Additional simulation results



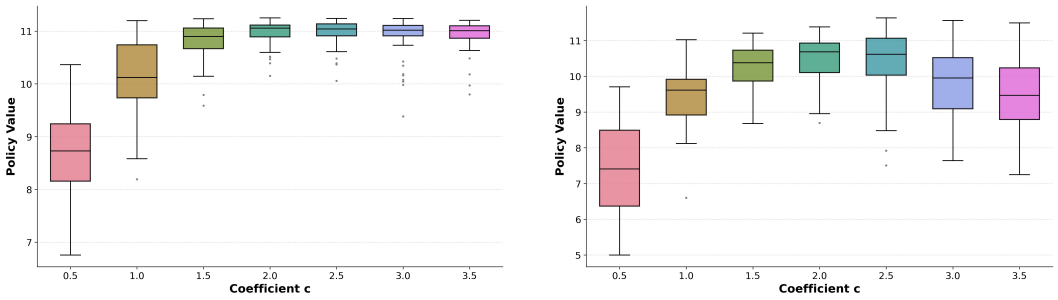
(a) Scenario 1 (b) Scenario 2

Figure 4: Boxplots of estimated values at sample size $n = 3000$.

D.4 Sensitivity analysis

In the numerical studies presented in Section 6, the bandwidth ϕ_n is selected as $\phi_n = c\hat{\sigma}_A n^{-1/6}$, where $c \in \mathcal{C} = \{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$. To evaluate the sensitivity of model performance to the choice of bandwidth, we implement $d_1(X, W)$ to estimate linear and nonlinear policies under Scenario 1 and Scenario 2, respectively, and report the results across the range of values for c.

As illustrated in Figure 5, the value function of our proposed estimator initially increases and then decreases, achieving its maximum at $c = 2$. This observation is consistent with the findings in Wu et al. (2023), which characterize the behavior of the estimated cMSE as a function of bandwidth.



(a) Scenario 1 (b) Scenario 2

Figure 5: Sensitivity analysis for $d_1(X, W)$ under Scenarios 1 and 2 with n = 2,000.

D.5 Additional scenarios with U-dependent optimal rules

We further consider two additional scenarios in which the oracle optimal dose depends on latent confounders. The data-generating process for (U, X, W, Z) remains the same as in Appendix D.1, while the treatment assignment and outcome mechanisms are modified as follows.

Scenario 3: The oracle optimal dose is a linear function of the confounders. The treatment is generated as

$$A = -U_1 + 0.8U_2 + \varepsilon_A, \qquad \varepsilon_A \sim \mathcal{N}(0, 1),$$

and the outcome Y is obtained by adding random noise $\varepsilon_Y \sim \mathcal{N}(0, 2^2)$ to

$$\mathbb{E}\left[Y \mid U, A, X, W, Z\right] = 4X_1 - 2X_2 - 2X_3 + 7U_2 + 10\exp\left\{-\frac{(A - d^*(W, U))^4}{0.1}\right\},$$

with the oracle optimal rule

$$d^*(W, U) = 1 + 0.5W_1 + 0.3U_2.$$

Scenario 4: The oracle optimal dose is a nonlinear function of the confounders. The treatment is generated as

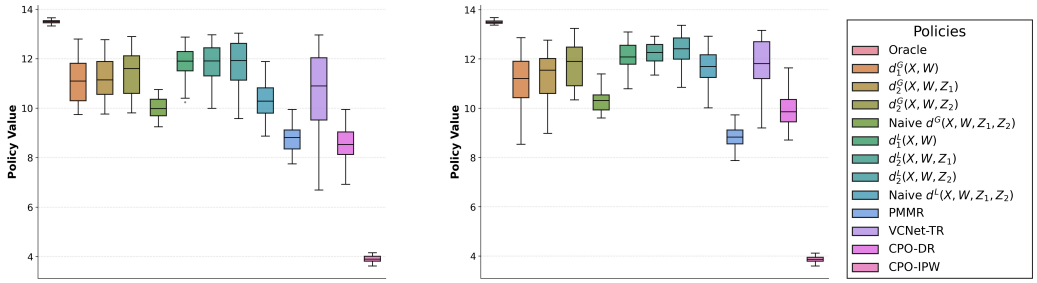
$$A = -U_1 + 0.8U_2 + \varepsilon_A, \qquad \varepsilon_A \sim \mathcal{N}(0, 1),$$

and the outcome Y is obtained by adding random noise $\varepsilon_Y \sim \mathcal{N}(0, 2^2)$ to

$$\mathbb{E}\left[Y \mid U, A, X, W, Z\right] = 2\cos(2\pi X_2) - 4X_3^3 + 7U_2 + 10\exp\left\{-\frac{(A - d^*(U))^4}{0.1}\right\},$$

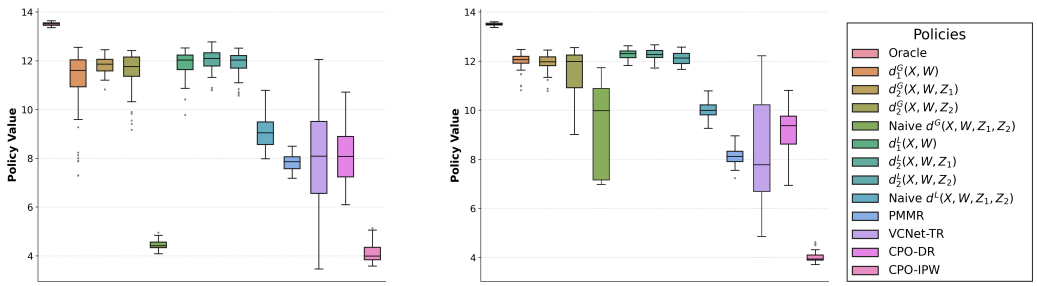
with the oracle optimal rule

$$d^*(U) = 0.7U_1^3 + 0.5\log(|U_2| + 1) - 1.$$



(a) sample size $n = 1500$ (b) sample size $n = 3000$

Figure 6: Boxplots of estimated values in Scenario 3.



(a) sample size $n = 1500$ (b) sample size $n = 3000$

Figure 7: Boxplots of estimated values in Scenario 4.

We implement the simulations as in Section 6. Figures 6 and 7 show the value functions of different estimators in Scenarios 3 and 4, respectively. These scenarios further assess the benefit of incorporating proxies into the decision rule. In particular, when the oracle optimal rule depends on latent confounders, allowing the decision rule to additionally depend on the proxy variables Z_1 or Z_2 enables d_2 to better capture information about the latent structure, leading to better performance than d_1. This is consistent with the intuition that the treatment-inducing proxies carry information about the latent confounders relevant to the optimal dose.

E Demand data generating process

The Demand experiment is generated from 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: $Y = A \times \min\{\exp((W - A)/10), 5\} - 5g(U) + \epsilon_5$.

Here, $g(u) = 2\left(\frac{(u-5)^4}{600} + e^{-4(u-5)^2} + \frac{u}{10} - 2\right)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$ independently for $i = 1, \ldots, 5$.

F Compute Resources

All experiments were run on a CPU-only server with 32 vCPUs (AMD EPYC 7T83 64-Core Processor), 60 GB of RAM, a 30 GB system disk, and a 50 GB data disk. The software environment used Ubuntu 22.04, Python 3.12, and PyTorch 2.5.1. Although the software image included CUDA 12.4, no GPU was allocated. We ran five threads in parallel. Each simulation scenario took approximately 3 hours, and the synthetic experiment took approximately 10 hours.