Conditional Counterfactual Mean Embeddings: Doubly Robust Estimation and Learning Rates

Thatchanon Anancharoenkij
Department of Statistics
Faculty of Science
Chiang Mai University
thatchanon_ananch@cmu.ac.th

Donlapark Ponnoprat
Department of Statistics
Faculty of Science
Chiang Mai University
donlapark.p@cmu.ac.th

Abstract

A complete understanding of heterogeneous treatment effects requires fully capturing the conditional distribution of the counterfactual outcomes. To this end, we study the *Conditional Counterfactual Mean Embeddings* (CCME), a framework that embeds conditional distributions of counterfactual outcomes into a reproducing kernel Hilbert space (RKHS). Under this framework, we develop a two-stage meta-estimator for CCME that accommodates any RKHS-valued regression in each stage. Based on this meta-estimator, we develop three practical CCME estimators: (1) *Ridge Regression* estimator, (2) *Deep Feature* estimator that parameterizes the feature map by a neural network, and (3) *Neural-Kernel* estimator that performs RKHS-valued regression, with the coefficients parameterized by a neural network. We provide finite-sample convergence rates for all estimators, establishing that they possess the double robustness property. Our experiments demonstrate that our estimators accurately recover distributional features including multimodal structure of conditional counterfactual distributions. Our code is available at this link.

1 Introduction

Causal inference is fundamental to scientific inquiry, policy design, and development of reliable artificial intelligence, where the primary goal is to evaluate the impact of interventions on outcomes of interest. Consider a scenario where a decision maker wants to understand the causal effect of a binary treatment $A \in \{0, 1\}$ on an outcome Y, with particular interest in how this effect varies across subpopulations characterized by covariates V. The potential outcomes framework [Rub74, Hol86] provides a principled approach to this setting: it posits that each unit has two *counterfactuals* Y^0 and Y^1. This framework highlights two fundamental challenges: (i) for each unit, the outcome under the alternative treatment remains unobserved, and (ii) a broader set of covariates X may confound the relationship between treatment and outcome. Given these challenges, the literature has traditionally focused on estimating scalar summaries, such as the conditional average treatment effect $\mathbb{E}[Y^1 - Y^0 \mid V = v]$.

While these scalar summaries provide valuable high-level insights, they often obscure critical nuances in the underlying distribution: different counterfactual distributions can share identical means yet exhibit vastly different shapes [KBW23]. A multimodal counterfactual distribution often indicates the presence of distinct underlying subgroups with different responses to the treatment. Capturing these behaviors calls for the estimation of the *conditional counterfactual distributions*, that is, the distributions of Y^0 and Y^1 conditional on $V = v$.

To study the conditional counterfactual distributions, we leverage the framework of *Kernel Mean Embeddings* (KME) [MFSS17], which allows us to manipulate probability distributions as vectors in a *reproducing kernel Hilbert space* (RKHS). This framework enables a wide range of statistical

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

tasks, such as density estimation [SZS+08, SFG+17, KF14], hypothesis testing [GBR+12, CSG16] and sampling [Wel09, CWS10]. For this reason, KME has seen considerable success as a tool for studying counterfactual distributions [MKSM21, PSSM21, SXG24].

These successes motivate us to study the *Conditional Counterfactual Mean Embeddings* (CCME), the embeddings of conditional counterfactual distributions in the RKHS, allowing us to extend the analysis of counterfactual distributions when the distributions vary across subpopulations. Inheriting the representation properties of the KME, the CCME enables a wide range of statistical tasks on conditional counterfactual distributions, such as density estimation, conditional two-sample testing, and sampling.

To estimate the CCME, we utilize *doubly robust* methods [RRZ94, SRR99]. The classical doubly robust methods debias the outcome variable by incorporating two models that describe the relationships between (i) the outcome and the covariates, and (ii) the treatment assignment and the covariates. These methods yield estimators that are consistent even if only one of the two models is correctly specified. The key insight of our approach is to extend these methods to the RKHS setting by instead *debiasing the feature map of the outcome*. As a result, we obtain CCME estimators in the functional space that inherit the double robustness property.

Contribution. We study the *Conditional Counterfactual Mean Embeddings* (CCME) as a way to represent conditional counterfactual distributions in an RKHS and develop doubly robust estimators through the double machine learning framework. Our specific contributions are three-fold:

1. **A Meta-Estimator for CCME:** We propose a two-stage meta-estimator for CCME that accommodates arbitrary propensity score models and conditional mean embedding (CME) estimators in the first stage, and any CME estimator in the second stage.
2. **Three Specific CCME Estimators:** We develop three practical instances of the meta-estimator, namely the *Ridge Regression* estimator, the *Deep Feature* estimator, and the *Neural-Kernel* estimator, each of which offers distinct advantages in terms of simplicity, flexibility and scalability.
3. **Finite-Sample Convergence Rates:** We establish finite-sample error bounds for all three estimators that explicitly track how convergence rates depend on the dimensions of the outcome, the covariates, and the conditioning variables. Crucially, we prove that our estimators achieve *double robustness in rates*, i.e., the first-stage error is controlled by the minimum of the rates achieved by the propensity score estimator and the CME estimator.

2 Background and Notations

We briefly describe the framework to study conditional counterfactual distributions. As a complement, we provide full definitions of positive definite kernels and Bochner integrals in Appendix H.

2.1 Problem Setup

We observe an i.i.d. sample $Z_1, \dots, Z_{2n}$ where $Z_i \coloneqq (X_i, A_i, Y_i) \sim P$. Here, $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is a vector of covariates, $A \in \{0, 1\}$ is a binary treatment assignment, and $Y \in \mathcal{Y} \subset \mathbb{R}^{d_y}$ is the observed outcome. We adopt the potential outcomes framework [Rub74, Hol86], under which $Y = AY^1 + (1-A)Y^0$, where Y^1 and Y^0 are referred to as the *potential outcomes* or *counterfactuals* under treatment ($A = 1$) and control ($A = 0$), respectively. Our focus is on the conditional distribution of the counterfactual Y^1 given covariates $V \coloneqq \eta(X) \in \mathcal{V} \subset \mathbb{R}^{d_v}$, where $\eta : \mathcal{X} \to \mathcal{V}$ is a known function (e.g., a projection onto a subset of covariates).

2.2 Reproducing Kernel Hilbert Spaces

Let $\mathcal{H}_\mathcal{Y}$ denote an RKHS on the outcome space $\mathcal{Y}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{Y}}$ and reproducing kernel $k_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The kernel is *reproducing* in the sense that $\langle g, k_\mathcal{Y}(\cdot, y) \rangle_{\mathcal{H}_\mathcal{Y}} = g(y)$ for any $g \in \mathcal{H}_\mathcal{Y}$ and $y \in \mathcal{Y}$. The *canonical feature map* $\phi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ is defined as $\phi(y) \coloneqq k_\mathcal{Y}(\cdot, y)$, which satisfies $k_\mathcal{Y}(y, y') = \langle \phi(y), \phi(y') \rangle_{\mathcal{H}_\mathcal{Y}}$.

The *mean embedding* of a random variable $Y \in \mathcal{Y}$ is defined as $\mu_Y \coloneqq \mathbb{E}_Y[\phi(Y)] \in \mathcal{H}_\mathcal{Y}$, where the expectation is the Bochner integral. When $k_\mathcal{Y}$ is a *characteristic kernel*, the mapping from the space

of probability distributions to the RKHS is injective [FGSS07, SGF+10]. This property ensures that μ_Y encodes all information about the distribution of Y.

The *conditional mean embedding* (CME) $\mu_{Y|X} : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ is defined as $\mu_{Y|X}(x) \coloneqq \mathbb{E}[\phi(Y) \mid X = x]$, which satisfies $\mathbb{E}[g(Y) \mid X = x] = \langle g, \mu_{Y|X}(x)\rangle_{\mathcal{H}_\mathcal{Y}}$ for any $g \in \mathcal{H}_\mathcal{Y}$.

2.3 The Target Estimand

Our goal is to estimate the conditional distribution of the counterfactual outcome Y^1 given a set of covariates $V = \eta(X)$. In the RKHS framework, we represent the conditional distribution by the *Conditional Counterfactual Mean Embeddings* (CCME):

$$\mu_{Y^1|V}(v) \coloneqq \mathbb{E}[\phi(Y^1) \mid V = v] \in \mathcal{H}_\mathcal{Y}.$$

This embedding fully characterizes the conditional distribution of Y^1 given $V = v$ when the kernel $k_\mathcal{Y}$ is *characteristic* [FGSS07, SGF+10] (e.g., Gaussian, Matérn, and Laplace kernels). As the conditional distribution of Y^1 can differ from that of Y, the embedding $\mu_{Y^1|V}(v)$ might not be the same as $\mu_{Y|V}(v) \coloneqq \mathbb{E}[\phi(Y) \mid V = v]$.

3 Doubly Robust Identification of CCME

To identify the CCME from observed data (X, A, Y), we adopt a standard technique for treatment effect estimation from observational data [RR83, Ros87] and extend it to RKHS settings.

First, we introduce two *nuisance functions* that model the conditional distributions of A and Y given covariates X: the *propensity score* $\pi(x) \coloneqq P(A = 1|X = x)$ and the CME of the treated outcome conditional on X, $\mu_0(x) \coloneqq \mathbb{E}[\phi(Y)|X = x, A = 1]$. Using these nuisance functions, we construct an RKHS-valued pseudo-outcome $\xi(Z) \in \mathcal{H}_\mathcal{Y}$ as:

$$\xi(Z) \coloneqq \frac{A}{\pi(X)}(\phi(Y) - \mu_0(X)) + \mu_0(X). \tag{1}$$

For identification of CCME, we require the following two standard assumptions:

Assumption 1. *(i) Conditional ignorability:* $Y^a \perp A \mid X$ *for each* $a \in \{0, 1\}$. *(ii) Strong positivity: There exists* $\varepsilon > 0$ *such that* $\varepsilon \leq \pi(x) \leq 1 - \varepsilon$ *for all* $x \in \mathcal{X}$.

Assumption 2. *The kernel* $k_\mathcal{Y}$ *is continuous and* $\sup_{y\in\mathcal{Y}} k_\mathcal{Y}(y, y) \leq B_k < \infty$ *for some* $B_k > 0$.

The first assumption is common in theoretical analysis of inverse propensity weighting and doubly robust estimators [RR83, KBW23]. Specifically, it ensures that treatment assignment is independent of potential outcomes given covariates X, i.e., there are no unmeasured confounders. Strong positivity requires that the propensity score is uniformly bounded away from 0 and 1, ensuring that both treatment groups are represented across all covariate values.

The boundedness of kernels is a mild assumption that is satisfied by commonly used kernels, e.g., Gaussian, Matérn, and Laplace kernels, but not by polynomial and linear kernels.

Under these assumptions, the CCME is identified via $\xi(Z)$. The proof is provided in Appendix I.

Proposition 1. *Under Assumptions 1 and 2, the CCME satisfies the following identification result:*

$$\mu_{Y^1|V}(v) = \mathbb{E}[\xi(Z) \mid V = v]. \tag{2}$$

4 The Meta-Estimator

The identification result from the previous section suggests a two-stage estimation strategy that has proven successful for estimating conditional treatment effects [CCD+18, SC21, Ken23]: first, estimate the nuisance functions π and μ_0; second, construct pseudo-outcomes $\widehat{\xi}(Z)$ and regress them on V to obtain the CCME estimator. Formally, we randomly split the full dataset $\mathcal{D}$ of size $2n$ into two disjoint folds, $\mathcal{D}_0$ and $\mathcal{D}_1$, each of size n. The two-stage estimation then proceeds as follows:

Stage 1: Nuisance Estimation. Using $\mathcal{D}_0$, we estimate two *nuisance functions*: (1) a propensity score estimator $\widehat{\pi}$ of π, and (2) a CME estimator $\widehat{\mu}_0$ of μ_0. The propensity score can be estimated

using any classification model that provides probability predictions. For the CME, several estimators have been proposed [SHSF09, XCS+21, SFS24], which we explore in the next section.

Stage 2: Pseudo-Outcome Regression. Using $\mathcal{D}_1$, we construct pseudo-outcomes from the estimated nuisance functions. For each $z = (x, a, y) \in \mathcal{X} \times \{0, 1\} \times \mathcal{Y}$, the pseudo-outcome $\widehat{\xi}(z) \in \mathcal{H}_\mathcal{Y}$ is constructed as:

$$\widehat{\xi}(z) \coloneqq \frac{a}{\widehat{\pi}(x)}(\phi(y) - \widehat{\mu}_0(x)) + \widehat{\mu}_0(x). \tag{3}$$

Given the observations $Z_{1i} \coloneqq (X_{1i}, A_{1i}, Y_{1i}) \in \mathcal{D}_1$ with the covariates $V_{1i} \coloneqq \eta(X_{1i})$, we estimate the CCME by regressing the pseudo-outcomes on the covariates. Specifically, we solve a regularized empirical risk minimization problem over a hypothesis class $\mathcal{F}$ of $\mathcal{H}_\mathcal{Y}$-valued functions:

$$\widehat{\mu}_{Y^1|V} \coloneqq \arg\min_{\mu \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \|\mu(V_{1i}) - \widehat{\xi}(Z_{1i})\|^2_{\mathcal{H}_\mathcal{Y}} + \Lambda(\mu), \tag{4}$$

where $\Lambda : \mathcal{F} \to \mathbb{R}_+$ is a regularization functional that controls model complexity. The choice of $\mathcal{F}$ and Λ determines the specific estimator. In the next section, we present three specific choices of $\mathcal{F}$ and Λ that yield practical estimators for the CCME.

5 Three Practical Estimators

In this section, we propose three choices of hypothesis class $\mathcal{F}$ for both first- and second-stage CME regression, yielding three computationally tractable estimators for the CCME with different computational-statistical trade-offs.

Throughout, we work with the two splits of data: $\mathcal{D}_0 \coloneqq \{Z_{0i}\}_{i=1}^n$ and $\mathcal{D}_1 \coloneqq \{Z_{1i}\}_{i=1}^n$, where $Z_{0i} \coloneqq (X_{0i}, A_{0i}, Y_{0i})$ and $Z_{1i} = (X_{1i}, A_{1i}, Y_{1i})$.

Let $\mathbf{I}_n$ be the $n \times n$ identity matrix and Id the identity operator on $\mathcal{H}_\mathcal{Y}$. Define an $\mathcal{H}_\mathcal{Y}$-valued vector $\widehat{\mathbf{\Xi}} \coloneqq (\widehat{\xi}(Z_{11}), \ldots, \widehat{\xi}(Z_{1n}))^\top$, which maps any $\mathbf{u} \in \mathbb{R}^n$ to $\mathcal{H}_\mathcal{Y}$ as follows: $\widehat{\mathbf{\Xi}}\mathbf{u} = \sum_{i=1}^n u_i \widehat{\xi}(Z_{1i})$.

5.1 Ridge Regression Estimator

We use the kernel ridge regression (KRR) [SHSF09] with V_{1i} as the input and $\widehat{\xi}(Z_{1i})$ as the target. To formalize this, we let $k_\mathcal{V} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a kernel on $\mathcal{V}$ with associated RKHS $\mathcal{H}_\mathcal{V}$. Denote by $\mathcal{BL}(\mathcal{H}_\mathcal{Y}; \mathcal{H}_\mathcal{Y})$ the set of bounded linear operators from $\mathcal{H}_\mathcal{Y}$ to itself. We define the operator-valued kernel $\Gamma_\mathcal{V} : \mathcal{V} \times \mathcal{V} \to \mathcal{BL}(\mathcal{H}_\mathcal{Y}; \mathcal{H}_\mathcal{Y})$ as $\Gamma_\mathcal{V}(v, v') \coloneqq k_\mathcal{V}(v, v')\,\mathrm{Id}$, and let $\mathcal{H}_{\Gamma_\mathcal{V}}$ be the associated vector-valued RKHS. Setting $\mathcal{F} = \mathcal{H}_{\Gamma_\mathcal{V}}$ in the least squares problem (4) with regularizer $\Lambda = \lambda_1 \|\cdot\|^2_{\mathcal{H}_{\Gamma_\mathcal{V}}}$ for some $\lambda_1 > 0$ yields the following problem:

$$\widehat{\mu}_{\mathrm{RR}} = \arg\min_{\mu \in \mathcal{H}_{\Gamma_\mathcal{V}}} \frac{1}{n} \sum_{i=1}^{n} \|\mu(V_{1i}) - \widehat{\xi}(Z_{1i})\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda_1 \|\mu\|^2_{\mathcal{H}_{\Gamma_\mathcal{V}}}.$$

The Ridge Regression estimator $\widehat{\mu}_{\mathrm{RR}}$ can be written explicitly as [SHSF09, GLG+12, PM20]:

$$\widehat{\mu}_{\mathrm{RR}}(v) = \widehat{\mathbf{\Xi}}(\mathbf{K}_V + n\lambda_1 \mathbf{I}_n)^{-1} \mathbf{k}_V(v), \tag{5}$$

where $\mathbf{K}_V \coloneqq (k_\mathcal{V}(V_{1i}, V_{1j}))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ and $\mathbf{k}_V(v) \coloneqq (k_\mathcal{V}(v, V_{11}), \ldots, k_\mathcal{V}(v, V_{1n}))^\top$.

5.2 Deep Feature Estimator

The empirical CCME estimator in (5) requires $O(n^3)$ time for matrix inversion, making it computationally expensive for large sample sizes. As an alternative, we adapt from [XCS+21] the *Deep Feature* estimator, which takes the form $\mu_{\mathrm{DF}}(v) = C\psi_\theta(v)$, where $\psi_\theta : \mathcal{V} \to \mathbb{R}^M$ is a neural network, parameterized by θ, that learns a feature map of $M \ll n$ dimensions, and $C : \mathbb{R}^M \to \mathcal{H}_\mathcal{Y}$ is a linear operator. This reduces the matrix inversion to $O(M^3)$ time.

To formalize the Deep Feature estimator, we denote by Θ a parameter space for neural networks from $\mathcal{V}$ to $\mathbb{R}^M$ and $\mathcal{BL}(\mathbb{R}^M; \mathcal{H}_\mathcal{Y})$ the set of bounded linear operators from $\mathbb{R}^M$ to $\mathcal{H}_\mathcal{Y}$. We then solve the

following variant of the least squares problem (4):

$$\underset{\substack{\theta\in\Theta,\\ C\in\mathcal{BL}(\mathbb{R}^M;\mathcal{H}_\mathcal{Y})}}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^{n}\|C\psi_\theta(V_{1i}) - \widehat{\xi}(Z_{1i})\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda_1\|C\|^2_{\text{HS}}\right), \tag{6}$$

where $\|\cdot\|_{\text{HS}}$ is the Hilbert-Schmidt norm and $\lambda_1 > 0$ is the regularization parameter. For each fixed ψ_θ, solving (6) with respect to C yields the following exact solution:

$$\widehat{C}_{\psi_\theta} = \widehat{\boldsymbol{\Xi}}\boldsymbol{\Psi}_\theta(\boldsymbol{\Psi}_\theta^\top\boldsymbol{\Psi}_\theta + n\lambda_1\mathbf{I}_M)^{-1}, \tag{7}$$

where $\boldsymbol{\Psi}_\theta \coloneqq (\psi_\theta(V_{11}),\ldots,\psi_\theta(V_{1n}))^\top \in \mathbb{R}^{n\times M}$. Substituting $C = \widehat{C}_{\psi_\theta}$ into (6) and dropping the term that does not depend on θ yield the following loss for ψ_θ:

$$\widehat{\mathcal{L}}_{\text{DF}}(\theta) = -\operatorname{Tr}\left(\mathbf{K}_{\widehat{\xi}}\boldsymbol{\Psi}_\theta(\boldsymbol{\Psi}_\theta^\top\boldsymbol{\Psi}_\theta + n\lambda_1\mathbf{I}_M)^{-1}\boldsymbol{\Psi}_\theta^\top\right), \tag{8}$$

where $\mathbf{K}_{\widehat{\xi}} \coloneqq (\langle\widehat{\xi}(Z_{1i}),\widehat{\xi}(Z_{1j})\rangle_{\mathcal{H}_\mathcal{Y}})_{i,j=1}^n \in \mathbb{R}^{n\times n}$.

Optimizing the loss with gradient descent, we obtain a learned feature map $\psi_{\widehat{\theta}}$. Our estimator is obtained by applying the optimal linear operator $\widehat{C}_{\psi_{\widehat{\theta}}}$ defined in (7):

$$\widehat{\mu}_{\text{DF}}(v) \coloneqq \widehat{C}_{\psi_{\widehat{\theta}}}\psi_{\widehat{\theta}}(v) = \widehat{\boldsymbol{\Xi}}\boldsymbol{\Psi}_{\widehat{\theta}}(\boldsymbol{\Psi}_{\widehat{\theta}}^\top\boldsymbol{\Psi}_{\widehat{\theta}} + n\lambda_1\mathbf{I}_M)^{-1}\psi_{\widehat{\theta}}(v).$$

5.3 Neural-Kernel Estimator

The Neural-Kernel estimator [SFS24] offers a third approach for CME estimation that completely eliminates the matrix inversion and the regularization parameter λ_1.

To formalize this estimator, we let $\widetilde{\mathcal{Y}}_M \coloneqq \{\widetilde{y}_j\}_{j=1}^M$ be a set of M grid points in $\mathcal{Y}$, which we specify before estimation.

Let $f_\theta : \mathcal{V} \to \mathbb{R}^M$ be a neural network parameterized by $\theta \in \Theta$ with M outputs, and $f_\theta(v)_j$ be the j-th component of $f_\theta(v)$. The Neural-Kernel estimator takes the form $\mu_{\text{NK}}(v) = \sum_{j=1}^M f_\theta(v)_j\phi(\widetilde{y}_j)$. We then solve the following variant of the least squares problem (4):

$$\underset{\theta\in\Theta}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{M} f_\theta(V_{1i})_j\phi(\widetilde{y}_j) - \widehat{\xi}(Z_{1i})\right\|^2_{\mathcal{H}_\mathcal{Y}}.$$

Expanding the squared norm using the reproducing property yields the following loss:

$$\widehat{\mathcal{L}}_{\text{NK}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[f_\theta(V_{1i})^\top\mathbf{K}_M f_\theta(V_{1i}) - 2f_\theta(V_{1i})^\top\mathbf{b}_i\right], \tag{9}$$

where $\mathbf{K}_M \coloneqq (k_\mathcal{Y}(\widetilde{y}_j,\widetilde{y}_l))_{j,l=1}^M$ and $\mathbf{b}_i \coloneqq (\langle\phi(\widetilde{y}_1),\widehat{\xi}(Z_{1i})\rangle_{\mathcal{H}_\mathcal{Y}},\ldots,\langle\phi(\widetilde{y}_M),\widehat{\xi}(Z_{1i})\rangle_{\mathcal{H}_\mathcal{Y}})^\top$.

Now suppose that the first-stage CME estimator $\widehat{\mu}_0$ is also a Neural-Kernel estimator *with the same grid points*. In other words, $\widehat{\mu}_0(x) = \sum_{j=1}^M g_{\widehat{\theta}_0}(x)_j\phi(\widetilde{y}_j)$ where $g_{\widehat{\theta}_0}$ is the learned network from the first stage. Plugging the pseudo-outcome formula from (3) into the definition of $\mathbf{b}_i$, we obtain:

$$\mathbf{b}_i = \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\mathbf{k}_i + \left(1 - \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\right)\mathbf{K}_M g_{\widehat{\theta}_0}(X_{1i}),$$

where $\mathbf{k}_i \coloneqq (k_\mathcal{Y}(\widetilde{y}_1,Y_{1i}),\ldots,k_\mathcal{Y}(\widetilde{y}_M,Y_{1i}))^\top$. Substituting this expression into (9) yields:

$$\widehat{\mathcal{L}}_{\text{NK}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[f_\theta(V_{1i})^\top\mathbf{K}_M f_\theta(V_{1i}) - 2\frac{A_{1i}}{\widehat{\pi}(X_{1i})}f_\theta(V_{1i})^\top\mathbf{k}_i - 2\left(1 - \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\right)f_\theta(V_{1i})^\top\mathbf{K}_M g_{\widehat{\theta}_0}(X_{1i})\right].$$

Hence, we can compute the loss in (9) solely in terms of the Gram matrix of the grid points and the neural network. Let $f_{\widehat{\theta}}$ be a neural network obtained by optimizing the loss with gradient descent. The final estimator is given by:

$$\widehat{\mu}_{\text{NK}}(v) \coloneqq \sum_{j=1}^{M} f_{\widehat{\theta}}(v)_j\phi(\widetilde{y}_j).$$

Choosing an estimator. The choice of estimator is mainly determined by sample size. The Ridge Regression estimator becomes computationally intractable for large n due to $O(n^3)$ matrix inversion costs, making the Deep Feature and Neural-Kernel estimators preferable in such regimes. Nonetheless, the Ridge Regression estimator has the advantages of having only two tuning parameters and generally smaller variance compared to the neural estimators.

Between the two neural estimators, the Deep Feature estimator typically exhibits lower variance due to the regularization, while the Neural-Kernel estimator can achieve comparable variance reduction through weight regularization or early stopping. For both methods, we recommend using cross-validation to select the strength of the regularization. We discuss related work in Appendix A.

6 Theoretical Analysis

We establish the convergence rates of our meta- and specific estimators. In preparation, we define the risks of the propensity score and the CME estimators as:

$$\mathcal{R}_\pi^2(\widehat{\pi}) := \mathbb{E}\big[(\widehat{\pi}(X) - \pi(X))^2\big], \qquad \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0) := \mathbb{E}\big[\|\widehat{\mu}_0(X) - \mu_0(X)\|_{\mathcal{H}_\mathcal{Y}}^2\big].$$

For the second stage, we define the risk of the $\mathcal{H}_\mathcal{Y}$-valued regression of the pseudo-outcome:

$$\mathcal{R}_{\widehat{\xi}}^2(\widehat{\mu}_{Y^1|V}) := \mathbb{E}\Big[\|\widehat{\mu}_{Y^1|V}(V) - \mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0]\|_{\mathcal{H}_\mathcal{Y}}^2\Big].$$

For all our bounds, we require the following assumption on the propensity score estimator analogous to Assumption 1, which can be achieved by truncating the propensity score estimator to $[\varepsilon, 1-\varepsilon]$:

Assumption 3. *The propensity score estimator satisfies $\varepsilon \leq \widehat{\pi}(x) \leq 1 - \varepsilon$ for all $x \in \mathcal{X}$.*

In the following result, we use $A \lesssim B$ and $A \asymp B$ when $A \leq cB$ and $c_1 B \leq A \leq c_2 B$, respectively, for some constants $c, c_1, c_2 > 0$ independent of n.

6.1 Meta-Estimator

We start with the convergence rate for the meta-estimator:

Theorem 1 (Meta-Estimator's Rate)**.** *Suppose that Assumptions 1, 2, and 3 hold. In addition, assume that $\sup_{x\in\mathcal{X}} \|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$ for some constant $B_{\mu_0} > 0$. Then the meta-estimator $\widehat{\mu}_{Y^1|V}$ from (4) satisfies:*

$$\mathbb{E}\big[\|\widehat{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim \mathcal{R}_{\widehat{\xi}}^2(\widehat{\mu}_{Y^1|V}) + \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}. \tag{10}$$

Here, the assumption $\sup_{x\in\mathcal{X}} \|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$ is mild, as we can control the norm of the CME estimator via regularization or truncation of the estimator's parameters.

The most important aspect of (10) is that the convergence rate of the meta-estimator is controlled by the rate of the second-stage estimator, as well as the *faster rate* of the two nuisance estimators, establishing the *double robustness* property of the meta-estimator.

Remark 1. *The last term on the right-hand side of (10) can also be replaced by a product of the two nuisance errors:*

$$\mathbb{E}\big[(\widehat{\pi}(X) - \pi(X))^4\big]^{1/2} \mathbb{E}\big[\|\widehat{\mu}_0(X) - \mu_0(X)\|_{\mathcal{H}_\mathcal{Y}}^4\big]^{1/2}, \tag{11}$$

as we show in Appendix O. We state (10) with the minimum because rates for nuisance estimators are usually given in L^2-norms, while passing from L^4- to L^2-norms requires the moment comparison condition of [FS23, Assumption 8]. The two bounds are not comparable in general. Under the moment comparison condition, the product is of smaller order than the minimum when both nuisance estimators converge, and of the same order when one of them does not. Without this condition, the L^4-norms can be much larger than the L^2-norms. In that case, the minimum can be the smaller bound.

6.2 Three Practical Estimators

In the following results, $W^{k,p}(\Omega)$ for any Euclidean set Ω denotes the Sobolev space of $L^p(\Omega)$ functions of order k (cf. Definition 3 in the Appendix), $\widehat{f}$ is the Fourier transform of a function f, $a \lesssim_{\log n} b$ if $a \leq c_1 (\log n)^{c_2} b$ for some constants $c_1, c_2 > 0$, and $\|\cdot\|$ is the standard Euclidean norm.

Here are the assumptions required for our theoretical analysis:

Assumption 4. *Y^1 and V are continuous random variables. The marginal distribution of V is strictly positive on $\mathcal{V}$ and has a density p_V with $\sup_{v\in\mathcal{V}} p_V(v) < \infty$. The domains $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{V} \subset \mathbb{R}^{d_v}$ are non-empty, bounded sets with Lipschitz boundaries and satisfy a uniform interior cone condition (see Definition 7).*

The assumption on Y^1 and V ensures the existence of the conditional density $\mathfrak{p}^1(y|v)$. The assumption on the domains is often required for analysis of Sobolev spaces (see, e.g., [Eva10, Chapter 5]). It is satisfied by any "well-behaved" domain that does not have any sharp cusps, such as balls or cubes.

Assumption 5. *The kernel $k_\mathcal{Y}$ is translation invariant. In other words, $k_\mathcal{Y}(y, y') = \varphi(y - y')$ for some function $\varphi : \mathbb{R}^{d_y} \to \mathbb{R}$. Its Fourier transform $\widehat{\varphi}(\omega)$ satisfies $\widehat{\varphi}(\omega) \lesssim (1 + \|\omega\|^2)^{-\tau}$, for some $\tau > d_y/2$. This implies that $\mathcal{H}_\mathcal{Y}$ is continuously embedded in the Sobolev space $W^{\tau,2}(\mathcal{Y})$.*

The translation invariance of the kernel allows us to analyze the smoothness of the kernel and the functions in the RKHS via their Fourier transforms (cf. Proposition 3). The Gaussian kernel, for example, satisfies this assumption for all $\tau > 0$. The Matérn kernel on $\mathbb{R}^{d_y}$ with smoothness parameter α satisfies the assumption with $\tau = \alpha + d_y/2$.

The next assumption concerns the smoothness of the conditional density function of Y^1 given $V = v$, denoted by $\mathfrak{p}^1(y|v)$:

Assumption 6. *For every outcome $y \in \mathcal{Y}$, the function $v \mapsto \mathfrak{p}^1(y|v)$ belongs to a Sobolev space $W^{r,2}(\mathcal{V})$ for some $r \in \mathbb{N}_0$ satisfying $r > d_v/2$. Moreover, $\int_\mathcal{Y} \|\mathfrak{p}^1(y|\cdot)\|^2_{W^{r,2}(\mathcal{V})} \, \mathrm{d}y < \infty$.*

Assumption 6 quantifies the smoothness of the conditional density $\mathfrak{p}^1$ in v via the Sobolev index r. As will be shown in our main results, r appears explicitly in our convergence rate bounds, indicating that densities that are smoother in v lead to faster rates.

6.2.1 Ridge Regression Estimator

As the second stage of the Ridge Regression estimator is the KRR estimator, we can use one of the existing results, e.g., [FS20, LMMG22, LMMG24], to establish the rate. In particular, as the estimator (5) is calculated based on the kernel function $k_\mathcal{V}$ on $\mathcal{V}$, we impose an assumption on this function instead of on $k_\mathcal{Y}$:

Assumption 5′. *The kernel of the conditioning variable $k_\mathcal{V}(v, v') = \varphi(v-v')$ is translation invariant. Its Fourier transform $\widehat{\varphi}(\omega)$ satisfies $\widehat{\varphi}(\omega) \asymp (1 + \|\omega\|^2)^{-\tau}$ for some $\tau > d_v/2$. This implies that $\mathcal{H}_\mathcal{V}$ is norm-equivalent to the Sobolev space $W^{\tau,2}(\mathcal{V})$.*

We now state our bound for the convergence rate of the Ridge Regression estimator:

Theorem 2. *Let the assumptions in Theorem 1 and Assumptions 4, 5′, and 6 hold with $\tau \geq r$. In addition, assume that $\inf_{v\in\mathcal{V}} p_V(v) > 0$.*

For the second-stage estimator $\widehat{\mu}_{\mathrm{RR}}$, by choosing $\lambda_1 \asymp n^{-\frac{\tau}{r+d_v/2}}$, the estimator achieves the bound:

$$\mathbb{E}\big[\|\widehat{\mu}_{\mathrm{RR}}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\big] \lesssim n^{-\frac{2r}{2r+d_v}} + \min\big\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\big\}.$$

6.2.2 Deep Feature Estimator

For both the Deep Feature and Neural-Kernel estimators, we still assume the translation invariance of $k_\mathcal{Y}$, with an additional assumption on the Fourier transform $\widehat{\varphi}$:

Assumption 7. *The kernel $k_\mathcal{Y}$ is translation invariant, $k_\mathcal{Y}(y, y') = \varphi(y - y')$, and, for any constant $C_y > 0$ such that $\mathcal{Y} \subset [-C_y, C_y]^{d_y}$, $\inf_{\|\omega\|\leq 26 d_y/C_y} \widehat{\varphi}(\omega) > 0$.*

As the RKHS-norm of the translation invariant kernel can be written as the L^2-norm weighted by the inverse of $\widehat{\varphi}$ (cf. Proposition 3), Assumption 7 ensures that $k_\mathcal{Y}$ is strictly positive definite, so that the Gram matrix of distinct points is invertible. This condition is satisfied by commonly used kernels, e.g., Gaussian, Matérn, and Laplace kernels, but not by polynomial or linear kernels.

Both estimators use fully connected ReLU networks with M outputs,

$$\psi_\theta(v) \coloneqq \mathbf{W}_L \, \sigma_{\mathrm{relu}}(\mathbf{W}_{L-1} \cdots \sigma_{\mathrm{relu}}(\mathbf{W}_1 v + \mathfrak{b}_1) \cdots + \mathfrak{b}_{L-1}) + \mathfrak{b}_L, \qquad \mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}, \ \mathfrak{b}_i \in \mathbb{R}^{d_i}, \tag{12}$$

where $\theta := (\mathbf{W}_i, \mathfrak{b}_i)_{i=1}^L \in \prod_{i=1}^L (\mathbb{R}^{d_i \times d_{i-1}} \times \mathbb{R}^{d_i})$ with $d_0 = d_v$ and $d_L = M$, and $\sigma_{\text{relu}}(x) := \max\{x, 0\}$ applied coordinatewise. Here, L is referred to as the *depth* and $d_1, \ldots, d_{L-1}$ the *widths* of the network. We write $W := \sum_{i=1}^L (d_i d_{i-1} + d_i)$ for the number of weights and biases.

With these notations, we present the convergence rates of the Deep Feature estimator, distinguishing between two cases based on the kernel type.

Theorem 3. *Let the assumptions in Theorem 1 and Assumptions 4, 6, and 7 hold. Let ψ_θ be a neural network in (12) with a total number of parameters W and depth L. In addition, assume that $\min_{1\le i\le L-1} d_i \asymp \max_{1\le i\le L-1} d_i$. Then, there exist constants $\bar{B} > 0$ and L_0, which only depend on r and d_v, such that, for any $B_{\mathcal{F}}^2 \ge \bar{B}$ and any $L \ge \max\{L_0, 3\}$, the Deep Feature estimator $\widehat{\mu}_{\text{DF}} = \widehat{C}_{\psi_{\widehat{\theta}}} \psi_{\widehat{\theta}}$ with the minimizing pair $(\widehat{\theta}, \widehat{C}_{\psi_{\widehat{\theta}}})$ of (6) over the set $\Theta := \{\theta : \sup_{v\in\mathcal{V}} \|\psi_\theta(v)\| \le B_{\mathcal{F}}\}$ satisfies the following statements with $b := 2r/d_v$:*

1. *If Assumption 5 holds, define $a := 2\tau/d_y$. By choosing $M \asymp n^{\frac{b}{4ab+4b+a}}$, $WL \asymp n^{\frac{a+2b}{4ab+4b+a}}$ and $\lambda_1 \asymp n^{-\frac{b(a+1)}{4ab+4b+a}}$, the estimator achieves*

$$\mathbb{E}[\|\widehat{\mu}_{\text{DF}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim_{\log n} n^{-\frac{ab}{4ab+4b+a}} + \min\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\}.$$

2. *If $k_\mathcal{Y}$ is a Gaussian kernel, by choosing $M \asymp (\log n)^{2d_y}$, $WL \asymp n^{\frac{1}{4b+1}}$ and $\lambda_1 \asymp n^{-\frac{b}{4b+1}}$, we have*

$$\mathbb{E}[\|\widehat{\mu}_{\text{DF}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim_{\log n} n^{-\frac{b}{4b+1}} + \min\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\}.$$

6.2.3 Neural-Kernel Estimator

For the Neural-Kernel estimator, we rely on an additional assumption:

Assumption 8. *The points $\{\widetilde{y}_j\}_{j=1}^M \subset \mathcal{Y}$ form a quasi-uniform grid with fill distance $h \asymp M^{-1/d_y}$, where the fill distance is defined as $h := \sup_{y\in\mathcal{Y}} \min_{1\le j\le M} \|y - \widetilde{y}_j\|$.*

Assumption 8 requires the grid points to be roughly uniformly spaced in $\mathcal{Y}$. Here is our result:

Theorem 4. *Let the assumptions in Theorem 1 and Assumptions 4, 6, 7, and 8 hold. Let f_θ be a neural network of the form (12) with a total number of parameters W and depth L. In addition, assume that $\min_{1\le i\le L-1} d_i \asymp \max_{1\le i\le L-1} d_i$. Then, there exist constants $\bar{B} > 0$ and L_0, which only depend on r and d_v, such that, for any $B_{\mathcal{F}}^2 \ge \bar{B}$ and any $L \ge \max\{L_0, 3\}$, the Neural-Kernel estimator $\widehat{\mu}_{\text{NK}}$ with the minimizer $\widehat{\theta}$ of (9) over the set $\Theta := \{\theta : \sup_{v\in\mathcal{V}} \|\sum_{j=1}^M f_\theta(v)_j \phi(\widetilde{y}_j)\|_{\mathcal{H}_\mathcal{Y}} \le B_{\mathcal{F}}\}$ satisfies the following statements with a and b as in Theorem 3.*

1. *If Assumption 5 holds, by choosing $M \asymp n^{\frac{b}{ab+a+2b}}$ and $WL \asymp n^{\frac{a+2b}{ab+a+2b}}$, we have*

$$\mathbb{E}[\|\widehat{\mu}_{\text{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim_{\log n} n^{-\frac{ab}{ab+a+2b}} + \min\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\}.$$

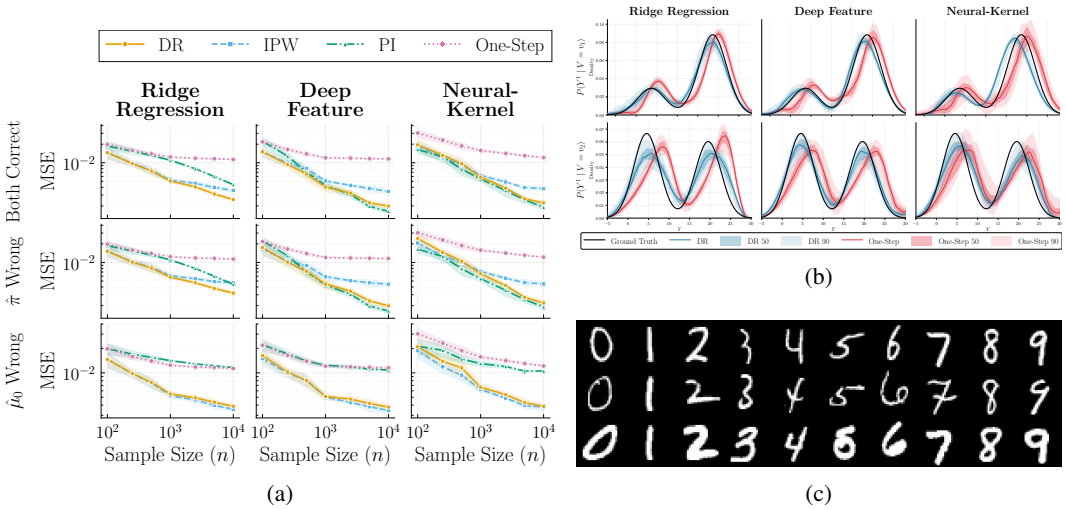
2. *If $k_\mathcal{Y}$ is a Gaussian kernel, by choosing $WL \asymp n^{\frac{1}{b+1}}$ and $M \asymp (\log n)^{2d_y}$, we have*

$$\mathbb{E}[\|\widehat{\mu}_{\text{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim_{\log n} n^{-\frac{2r}{2r+d_v}} + \min\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\}.$$

Discussion of the rates. For all estimators, the bounds improve with the smoothness r of $v \mapsto \mathfrak{p}^1(y|v)$ but deteriorate with the data dimensions. For the Deep Feature and Neural-Kernel estimators, the bounds also improve with the kernel smoothness τ. The bound of Theorem 3 decreases more slowly than the bound of Theorem 4. This difference arises because the Neural-Kernel estimator fits coefficients in a subspace of $\mathcal{H}_\mathcal{Y}$ that is fixed before the data are seen, whereas the Deep Feature estimator also learns the map C from the data, which contributes to the variance of the estimator.

The term $\min\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\}$ indicates that the first-stage error is governed by the nuisance with the smaller error. If $d_x > d_v$ and the first-stage CME estimator is of the same type as the second-stage estimator, the first-stage error will converge more slowly and become the bottleneck.

Remark 2. *The Deep Feature estimator and Neural-Kernel estimator were originally introduced in [XCS+21] and [SFS24] as CME estimators without finite-sample convergence rates. For the Deep*



(a) (b) (c)

Figure 1. (a) MSE vs. sample size for three methods (columns) under three misspecification scenarios (rows). Shaded regions indicate standard errors over 10 seeds; (b) Estimated counterfactual densities for profiles v_1 (top) and v_2 (bottom) defined in Appendix F over 30 runs at $n = 10000$. Blue: DR; Red: One-Step. Solid lines are medians; dark and light regions show pointwise 50th and 90th percentile intervals; (c) Empirical mode for each class under three density estimates. Row 1: Oracle estimates on clean data. Row 2: DR estimates (proposed). Row 3: One-Step estimates.

Feature estimator, [XCS+21, Appendix B] provides a bound in terms of the Rademacher complexity of the deep feature class, for a regression with values in a finite-dimensional feature space. In contrast, we evaluate the complexity of the network class for the $\mathcal{H}_\mathcal{Y}$-valued regression, which, combined with the universal approximation bound of [Yan25], yields a convergence rate in terms of the sample size. Consequently, the above theorems with $\widehat{\pi} = \pi$ and $\widehat{\mu}_0 = \mu_0$ establish the first known finite-sample convergence rates for these methods.

7 Experiments

To benchmark $\widehat{\mu}_\bullet(v)$ where $\bullet \in \{\mathrm{RR}, \mathrm{DF}, \mathrm{NK}\}$, we choose the task of estimation of conditional density $\mathfrak{p}^1(y|v)$ using $\widehat{\mathfrak{p}}^1_\bullet(y|v) := \langle \widehat{\mu}_\bullet(v), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}}$ with the Gaussian kernel $k_\mathcal{Y}(y, y') = (\sqrt{2\pi}\sigma)^{-d_y} \exp(-\|y - y'\|^2/(2\sigma^2))$. We note that $\widehat{\mathfrak{p}}^1_\bullet(y|v)$ is a biased estimate of $\mathfrak{p}^1(y|v)$, where the bias is $O(\sigma^2)$ (cf. Corollary 1 in the Appendix). The corresponding density estimators are provided in Appendix C. For each estimator, we compare four pseudo-outcome constructions: (i) the proposed *Doubly Robust* (DR) pseudo-outcome $\widehat{\xi}(Z)$ from (3); (ii) *Inverse Propensity Weighting* (IPW), $\widehat{\xi}_{\mathrm{IPW}}(Z) := \frac{A}{\widehat{\pi}(X)}\phi(Y)$; (iii) the *Plug-in* (PI), $\widehat{\xi}_{\mathrm{PI}}(Z) := \widehat{\mu}_0(X)$; and (iv) the *One-Step* estimator that directly uses the corresponding CME estimator with V as an input to estimate the CCME [PSSM21]; this is a conditional version of the plug-in estimator of [MKSM21]. Full experiment details are provided in Appendix F. Further applications of our DR estimator are demonstrated in Appendix G.

7.1 Analysis of Double Robustness

Setup. We generate covariates $X \sim \mathcal{N}(\mathbf{1}_{10}, \mathbf{I}_{10}) \in \mathbb{R}^{10}$ with the conditioning variable $V = (X_1, X_2, X_3, X_4, X_5)$. The propensity score is $\pi(X) = 0.1 + 0.8 \cdot \mathbb{I}(X_1 \in [0, 2] \text{ and } X_6 \geq 1.5)$. Our estimator for π is a random forest. We generate the outcomes according to $Y = 1 + X^\top\beta + A \cdot (2 + X^\top\gamma) + \epsilon$ where β and γ are fixed coefficients and ϵ is a heteroscedastic bimodal noise.

We consider three misspecification scenarios: (a) both nuisance models correctly specified, (b) π misspecified by replacing the random forest with a logistic regression, and (c) μ_0 misspecified by excluding X_6 from the outcome model. We vary $n \in \{100, 250, 500, 1000, 2500, 5000, 10000\}$ across 10 seeds and measure the MSE against the analytically computed ground truth $\mathfrak{p}^1(y|v)$.

Results. Figure 1a shows the convergence behavior. When both models are correct (top row), DR consistently achieves lower MSE than One-Step across all sample sizes and all three methods. This gap persists even at large sample sizes, indicating that One-Step suffers from selection bias. The DR, IPW, and PI variants of the estimators achieve comparable MSE in this setting; however, for the Ridge Regression estimator, PI exhibits slightly higher MSE than DR and IPW due to selection bias. For the Deep Feature and Neural-Kernel estimators, IPW performs worse than DR and PI due to the first-stage CME estimator.

When π is misspecified (middle row), IPW variants of the Deep Feature and Neural-Kernel estimators converge more slowly than DR and PI, as the latter two leverage the correct outcome model. Conversely, when μ_0 is misspecified (bottom row), PI variants of all estimators fail to converge while DR and IPW remain consistent via the correct propensity score. In all cases, DR matches the superior performance of either PI or IPW, confirming the double robustness property of our estimators.

To demonstrate our estimators' capacity for bias correction, we compare their density estimates for Y^1 against those from the corresponding One-Step estimators. The plots of conditional densities with two different values of V are shown in Figure 1b. We observe that the One-Step estimates show noticeable bias, especially in the Ridge Regression and Deep Feature estimators. The DR estimates, on the other hand, are closer to the ground truth, suggesting that the DR estimator effectively mitigates selection bias.

7.2 Semi-Synthetic MNIST Experiment

Setup. To evaluate scalability to high-dimensional outcomes, we consider a task of estimation of class-conditional densities from noisy MNIST images $\mathcal{D}_{\text{MNIST}} := \{Y_i\}_{i=1}^{60000}$, where $Y_i \in [0,1]^{784}$. One of our covariates is the image intensities $I := \{I_i\}_{i=1}^{60000}$, where $I_i := \frac{1}{784}\sum_{j=1}^{784} Y_{ij}$. Due to budget constraints, the denoiser prefers to denoise easy images that have high intensities. Thus, we can assign a treatment variable A_i to each image, with $A_i = 1$ if the image is clean and $A_i = 0$ if it is noisy. To simulate this task, we assume that the probability of an image being denoised is $\pi(I_i) = \mathsf{sigmoid}(3(I_i - q_{0.9}(I))/\mathsf{std}(I))$, where $\mathsf{sigmoid}(x) := 1/(1+e^{-x})$ and $q_{0.9}(I)$ is the 0.9 quantile of I, so that high-intensity images are more likely to be denoised. We let $Y_i^1 = Y_i$ and $\widetilde{Y}_i := Y_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 0.16\mathbf{I}_{784})$. Let V_i be the digit label and $X_i = (I_i, V_i)$ the covariate. Our goal is to recover the conditional densities of Y^1: $\mathfrak{p}^1(y|v)$ for each $v \in \{0,\dots,9\}$. We use the Deep Feature estimator in both stages, with a ReLU network of two hidden layers of 100 units and a Gaussian kernel with bandwidth 1. The remaining configuration is given in Table 2.

Results. Figure 1c shows the retrieved empirical modes $\widehat{y}_{\text{mode}}(v) := \arg\max_{y \in \mathcal{D}_{\text{MNIST}}} \widehat{\mathfrak{p}}^1(y|v)$ across all 10 classes. The DR estimator produces similar images to the oracle estimator, while the One-Step estimator retrieves the correct digits but with bolder strokes due to the selection bias toward high-intensity images. The DR estimator addresses this bias through the propensity score weighting.

8 Conclusion

We study the Conditional Counterfactual Mean Embeddings, a framework for representing conditional counterfactual distributions in an RKHS. We develop a two-stage meta-estimator based on doubly robust pseudo-outcomes and instantiate it with three estimators. Our analysis establishes convergence rates that explicitly track the dependence on the data dimensions and the smoothness of the conditional density, and shows that all three estimators achieve double robustness in rates. Experiments on synthetic and semi-synthetic data confirm that the DR estimators recover multimodal conditional counterfactual densities and correct for selection bias. A limitation is the lack of a lower bound for CCME estimation, which we plan to explore in future work. Moreover, the conditional distribution of $Y^1 - Y^0$ given V depends on the joint distribution of (Y^1, Y^0), which is identified only under additional assumptions such as rank invariance [HSC97]. Our estimators instead target the conditional distributions of Y^1 and Y^0 given V. These estimates can be used for functionals of the two distributions, such as quantile and conditional value-at-risk (CVaR) treatment effects (Appendix G).

Acknowledgements

TA and DP were financially supported by the Faculty of Science, Chiang Mai University, Thailand. TA also received the TA/RA Scholarship from Chiang Mai University.

References

[AB99] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[Aba02] Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292, 2002.

[ALdST07] Rémi Arcangéli, María Cruz López de Silanes, and Juan José Torrens. An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211, 2007.

[BFH+18] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. http://github.com/jax-ml/jax, 2018. Version 0.3.13.

[BHLM19] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[BK01] Peter J. Bickel and Jaimyoung Kwon. Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11(4):863–886, 2001.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BLO12] Francis R. Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, 2012.

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[CFVM13] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[CSG16] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615. PMLR, 2016.

[CWHM15] S. N. Chandler-Wilde, D. P. Hewett, and A. Moiola. Interpolation of Hilbert and Sobolev spaces: quantitative estimates and counterexamples. *Mathematika*, 61(2):414–443, 2015.

[CWS10] Yutian Chen, Max Welling, and Alexander J. Smola. Super-samples from kernel herding. In Peter Grünwald and Peter Spirtes, editors, *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8–11, 2010*, pages 109–116. AUAI Press, 2010.

[DFL96] John DiNardo, Nicole M. Fortin, and Thomas Lemieux. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044, 1996.

[Día17] Iván Díaz. Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference*, 190:39–51, 2017.

[Eva10] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, R.I., second edition, 2010.

[FGSS07] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3–6, 2007*, pages 489–496. Curran Associates, Inc., 2007.

[FHES24] Jake Fawkes, Robert Hu, Robin J. Evans, and Dino Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects. *Transactions on Machine Learning Research*, 2024.

[FS20] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

[FS23] Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[GBR+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[GLG+12] Steffen Grünewälder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, 2012.

[GM17] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1292–1301. PMLR, 2017.

[Hau95] David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[HDS22] Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient — a measure of conditional dependence. *Journal of Machine Learning Research*, 23(216):1–58, 2022.

[Hol86] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[HSC97] James J. Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

[HvNVW16] Tuomas Hytönen, Jan van Neerven, Mark Veraar, and Lutz Weis. *Analysis in Banach Spaces: Volume I: Martingales and Littlewood-Paley Theory*. Springer International Publishing, 2016.

[KBW23] Edward H. Kennedy, Sivaraman Balakrishnan, and Larry A. Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

[Ken23] Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[KF14] Motonobu Kanagawa and Kenji Fukumizu. Recovering distributions from Gaussian RKHS embeddings. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 457–465, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

[KG21] Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

[KKK24] Kwangho Kim, Jisu Kim, and Edward H. Kennedy. Causal effects based on distributional distances. *CoRR*, abs/1806.02935, 2024.

[KO23] Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6037–6060. PMLR, 25–27 Apr 2023.

[KR24] Ilmun Kim and Aaditya Ramdas. Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683–711, 2024.

[LC24] Alex Luedtke and Incheoul Chung. One-step estimation of differentiable Hilbert-valued parameters. *The Annals of Statistics*, 52(4):1534–1563, 2024.

[LLJ16] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284. PMLR, 2016.

[LMMG22] Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4433–4445. Curran Associates, Inc., 2022.

[LMMG24] Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Towards optimal Sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25(181):1–51, 2024.

[MFF23] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Normalizing flows for interventional density estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24361–24397. PMLR, 2023.

[MFSS17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017.

[MK24] Diego Martinez-Taboada and Edward Kennedy. Counterfactual density estimation using kernel Stein discrepancies. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net, 2024.

[MKSM21] Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

[MTRK23] Diego Martinez-Taboada, Aaditya Ramdas, and Edward Kennedy. An efficient doubly-robust test for the kernel treatment effect. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 59924–59952. Curran Associates, Inc., 2023.

[PM20] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.

[PSSM21] Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8401–8412. PMLR, 2021.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RM15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538. PMLR, 2015.

[Ros87] Paul R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[RR83] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RRZ94] 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.

[Rub74] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[RZ10] Christian Rieger and Barbara Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32(1):103–129, 2010.

[SC21] Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

[SFG+17] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

[SFS24] Eiki Shimizu, Kenji Fukumizu, and Dino Sejdinovic. Neural-kernel conditional mean embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45040–45059. PMLR, 2024.

[SGF+10] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

[SHSF09] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning*, pages 961–968. ACM, 2009.

[SMR22] Meyer Scetbon, Laurent Meunier, and Yaniv Romano. An asymptotic test for conditional independence using analytic kernel embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19328–19346. PMLR, 2022.

[SRR99] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[Ste70] Elias M. Stein. *Singular Integrals and Differentiability Properties of Functions*, volume 30 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.

[SXG24] Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516, 2024.

[SZS+08] Le Song, Xinhua Zhang, Alexander J. Smola, Arthur Gretton, and Bernhard Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning*, pages 992–999. ACM, 2008.

[TVE10] Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[vdVW23] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, Cham, second edition, 2023.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[Wel09] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1121–1128. ACM, 2009.

[Wen04] Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

[XCS+21] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net, 2021.

[Yan25] Yunfei Yang. On the optimal approximation of Sobolev and Besov functions using deep ReLU neural networks. *Applied and Computational Harmonic Analysis*, 79:101797, 2025.

[ZBG25] Houssam Zenati, Bariscan Bozkurt, and Arthur Gretton. Doubly-robust estimation of counterfactual policy mean embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[ZPJS11] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Fábio Gagliardi Cozman and Avi Pfeffer, editors, *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14–17, 2011*, pages 804–813. AUAI Press, 2011.

A Related Work

The estimation of counterfactual distributions has evolved through several approaches. Early work focused on estimating cumulative distribution functions under linear parametric assumptions [Aba02, CFVM13, Día17], while another line of work focused on estimating probability density functions using either kernel smoothing combined with propensity score weighting [DFL96] or doubly robust methods [BK01, KKK24]. More recently, [KBW23] proposed an alternative approach using cosine series approximation within a doubly robust framework. Other recent advances in density estimation include doubly robust methods via kernel Stein discrepancies [LLJ16, CSG16, GM17] developed by [MK24], and via normalizing flows [TVE10, RM15] developed by [MFF23].

A significant development occurred with the introduction of kernel mean embeddings (KME) into causal inference. [MKSM21] proposed Counterfactual Mean Embeddings, which represent *marginal* counterfactual distributions in an RKHS. They proposed to estimate these embeddings using the plug-in estimator, that is, the kernel ridge regression [SHSF09, GLG+12] on each treatment group. These embeddings enabled applications in treatment effect measurement and hypothesis testing of distributional effects; for example, doubly robust estimators for Counterfactual Mean Embeddings were employed to perform hypothesis testing of causal effects [MTRK23, FHES24].

Extensions to more general settings have also been explored. [LC24] established a comprehensive framework for one-step estimation of differentiable functional-valued parameters, proving consistency and asymptotic linearity of their estimator. [ZBG25] studied RKHS embeddings of counterfactual distributions under stochastic discrete interventions, which include binary treatments as a special case. They leveraged the results of [LC24] to propose a doubly robust estimator. However, a key limitation of both [LC24] and [ZBG25] is their focus on *marginal* counterfactual distributions rather than conditional counterfactual distributions.

The literature on *conditional* counterfactual distributions (CCD) remains relatively sparse. [PSSM21] introduced CCME for hypothesis testing of CCD, though their estimator is a one-step plug-in estimator, which may suffer from selection bias. In contrast, our doubly robust estimators correct this bias through propensity score weighting. [SXG24] developed kernel estimators of dose, heterogeneous, and incremental response functions, which include embeddings of conditional counterfactual distributions in continuous treatment regimes, whereas our work focuses on binary treatments. In a complementary direction, [KO23] developed doubly robust methods for estimating treatment effects on various statistics of CCD, including quantiles and conditional value-at-risk. However, our CCME framework offers a distinct advantage: by preserving the full distributional information as a functional object in an RKHS, it enables a broader range of downstream tasks such as density estimation [KF14], conditional independence testing [FGSS07, ZPJS11, HDS22, SMR22], and sampling [CWS10].

B Algorithms

The algorithms for the Ridge Regression estimator, Deep Feature estimator, and Neural-Kernel estimator are detailed in Algorithm 1, Algorithm 2, and Algorithm 3, respectively. In each algorithm, m denotes the number of treated units in $\mathcal{D}_0$.

Algorithm 1 Ridge Regression Estimator

Require: Dataset $\mathcal{D} \coloneqq \{Z_i\}_{i=1}^{2n}$ where $Z_i = (X_i, A_i, Y_i)$
Require: Kernels $k_{\mathcal{X}}, k_{\mathcal{V}}, k_{\mathcal{Y}}$ on $\mathcal{X}, \mathcal{V}, \mathcal{Y}$
Require: Propensity score estimation algorithm Alg_π
Require: Regularization parameters $\lambda_0, \lambda_1 > 0$

Randomly split $\mathcal{D}$ into $\mathcal{D}_0 = \{Z_{0i}\}_{i=1}^n$ and $\mathcal{D}_1 = \{Z_{1i}\}_{i=1}^n$
$\{Z_{0(i)}\}_{i=1}^m \leftarrow \{Z_{0i} \in \mathcal{D}_0 \mid A_{0i} = 1\}$
Stage 1: Nuisance Estimation on $\mathcal{D}_0$
  $\widehat{\pi} \leftarrow \mathsf{Alg}_\pi(\mathcal{D}_0)$
  Compute Gram matrix $\mathbf{K}_{X_0} \in \mathbb{R}^{m \times m}$ with $(\mathbf{K}_{X_0})_{ij} \coloneqq k_{\mathcal{X}}(X_{0(i)}, X_{0(j)})$
  $\mathbf{k}_{X_0}(x) \leftarrow (k_{\mathcal{X}}(x, X_{0(1)}), \ldots, k_{\mathcal{X}}(x, X_{0(m)}))^\top$
  $\widehat{\mu}_0(x) \leftarrow \sum_{j=1}^m \alpha_j(x)\phi(Y_{0(j)})$ with $\boldsymbol{\alpha}(x) \coloneqq (\mathbf{K}_{X_0} + m\lambda_0 \mathbf{I}_m)^{-1}\mathbf{k}_{X_0}(x)$
Stage 2: Pseudo-Outcome Regression on $\mathcal{D}_1$
  for $i = 1$ to n **do**
    $\widehat{\xi}(Z_{1i}) \leftarrow \frac{A_{1i}}{\widehat{\pi}(X_{1i})}(\phi(Y_{1i}) - \widehat{\mu}_0(X_{1i})) + \widehat{\mu}_0(X_{1i})$
  end for
  Compute Gram matrix $\mathbf{K}_V \in \mathbb{R}^{n \times n}$ with $(\mathbf{K}_V)_{ij} = k_{\mathcal{V}}(V_{1i}, V_{1j})$
  $\boldsymbol{\beta}(v) \leftarrow (\mathbf{K}_V + n\lambda_1 \mathbf{I}_n)^{-1}\mathbf{k}_V(v)$
  $\widehat{\mu}_{Y^1|V}(v) \leftarrow \sum_{i=1}^n \beta_i(v)\widehat{\xi}(Z_{1i})$
return $\widehat{\mu}_{Y^1|V}$

C Application to Conditional Counterfactual Density Estimation

The CCME can be used to recover the full conditional counterfactual density (CCD) $\mathfrak{p}^1(y|v)$. If $k_{\mathcal{Y}}$ is a translation-invariant kernel that integrates to one (e.g., the Gaussian kernel $k_{\mathcal{Y}}(y, y') = (\sqrt{2\pi}\sigma)^{-d_y} \exp(-\|y - y'\|^2/(2\sigma^2))$), the CCME satisfies:

$$\langle \mu_{Y^1|V}(v), \phi(y) \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{Y^1 \sim \mathfrak{p}^1(\cdot|v)}[k_{\mathcal{Y}}(Y^1, y)],$$

which is the kernel-smoothed density $\bar{\mathfrak{p}}^1(y|v)$. Thus, we obtain the estimated density $\widehat{\mathfrak{p}}^1(y|v)$ at any query point $y \in \mathcal{Y}$ by computing the inner product:

$$\widehat{\mathfrak{p}}^1(y|v) = \langle \widehat{\mu}_{Y^1|V}(v), \phi(y) \rangle_{\mathcal{H}_{\mathcal{Y}}}.$$

Algorithm 2 Deep Feature Estimator

Require: Dataset $\mathcal{D} = \{Z_i\}_{i=1}^{2n}$ where $Z_i = (X_i, A_i, Y_i)$
Require: Kernel $k_{\mathcal{Y}}$ on $\mathcal{Y}$
Require: Propensity score estimation algorithm Alg_π
Require: Neural network parameter spaces Θ_0, Θ_1
Require: Feature dimension M
Require: Regularization parameters $\lambda_0, \lambda_1 > 0$

Randomly split $\mathcal{D}$ into $\mathcal{D}_0 = \{Z_{0i}\}_{i=1}^n$ and $\mathcal{D}_1 = \{Z_{1i}\}_{i=1}^n$
$\{Z_{0(i)}\}_{i=1}^m \leftarrow \{Z_{0i} \in \mathcal{D}_0 \mid A_{0i} = 1\}$
Stage 1: Nuisance Estimation on $\mathcal{D}_0$
  $\widehat{\pi} \leftarrow \mathsf{Alg}_\pi(\mathcal{D}_0)$
  Initialize neural network $\psi_{0,\theta} : \mathcal{X} \to \mathbb{R}^M$ with $\theta \in \Theta_0$
  $\mathbf{\Psi}_{0,\theta} \leftarrow (\psi_{0,\theta}(X_{0(1)}), \ldots, \psi_{0,\theta}(X_{0(m)}))^\top$
  Compute Gram matrix $\mathbf{K}_{Y_0} \in \mathbb{R}^{m\times m}$ with $(\mathbf{K}_{Y_0})_{ij} \coloneqq k_{\mathcal{Y}}(Y_{0(i)}, Y_{0(j)})$
  $\mathcal{L}_0(\theta) \leftarrow -\mathrm{Tr}\big(\mathbf{K}_{Y_0}\mathbf{\Psi}_{0,\theta}(\mathbf{\Psi}_{0,\theta}^\top\mathbf{\Psi}_{0,\theta} + m\lambda_0\mathbf{I}_M)^{-1}\mathbf{\Psi}_{0,\theta}^\top\big)$
  $\widehat{\theta}_0 \leftarrow \arg\min_{\theta\in\Theta_0} \mathcal{L}_0(\theta)$
  $\widehat{\mathbf{\Psi}}_0 \leftarrow (\psi_{0,\widehat{\theta}_0}(X_{0(1)}), \ldots, \psi_{0,\widehat{\theta}_0}(X_{0(m)}))^\top$
Stage 2: Pseudo-Outcome Regression on $\mathcal{D}_1$
  Initialize neural network $\psi_\theta : \mathcal{V} \to \mathbb{R}^M$ with $\theta \in \Theta_1$
  $\mathbf{\Psi}_\theta \leftarrow (\psi_\theta(V_{11}), \ldots, \psi_\theta(V_{1n}))^\top$
  $\widehat{\mathbf{\Psi}}_{01} \leftarrow (\psi_{0,\widehat{\theta}_0}(X_{11}), \ldots, \psi_{0,\widehat{\theta}_0}(X_{1n}))$
  Compute Gram matrix $\mathbf{K}_{Y_1} \in \mathbb{R}^{n\times n}$ with $(\mathbf{K}_{Y_1})_{ij} \coloneqq k_{\mathcal{Y}}(Y_{1i}, Y_{1j})$
  Compute Gram matrix $\mathbf{K}_{Y_0Y_1} \in \mathbb{R}^{m\times n}$ with $(\mathbf{K}_{Y_0Y_1})_{ij} \coloneqq k_{\mathcal{Y}}(Y_{0(i)}, Y_{1j})$
  $\mathbf{C} \leftarrow \widehat{\mathbf{\Psi}}_0(\widehat{\mathbf{\Psi}}_0^\top\widehat{\mathbf{\Psi}}_0 + m\lambda_0\mathbf{I}_M)^{-1}\widehat{\mathbf{\Psi}}_{01}$
  $\mathbf{D}_\omega \leftarrow \mathrm{diag}(A_{11}/\widehat{\pi}(X_{11}), \ldots, A_{1n}/\widehat{\pi}(X_{1n}))$
  $\mathbf{D}_{1-\omega} \leftarrow \mathbf{I}_n - \mathbf{D}_\omega$
  $\mathbf{K}_{\widehat{\xi}} \leftarrow \mathbf{D}_\omega\mathbf{K}_{Y_1}\mathbf{D}_\omega + \mathbf{D}_\omega\mathbf{K}_{Y_0Y_1}^\top\mathbf{C}\mathbf{D}_{1-\omega} + \mathbf{D}_{1-\omega}\mathbf{C}^\top\mathbf{K}_{Y_0Y_1}\mathbf{D}_\omega + \mathbf{D}_{1-\omega}\mathbf{C}^\top\mathbf{K}_{Y_0}\mathbf{C}\mathbf{D}_{1-\omega}$
  $\mathcal{L}_1(\theta) \leftarrow -\mathrm{Tr}\Big(\mathbf{K}_{\widehat{\xi}}\mathbf{\Psi}_\theta(\mathbf{\Psi}_\theta^\top\mathbf{\Psi}_\theta + n\lambda_1\mathbf{I}_M)^{-1}\mathbf{\Psi}_\theta^\top\Big)$
  $\widehat{\theta}_1 \leftarrow \arg\min_{\theta\in\Theta_1} \mathcal{L}_1(\theta)$
  $\widehat{\mathbf{\Psi}}_1 \leftarrow (\psi_{\widehat{\theta}_1}(V_{11}), \ldots, \psi_{\widehat{\theta}_1}(V_{1n}))^\top$
  $\widehat{\mu}_{Y^1|V}(v) \leftarrow \sum_{i=1}^n c_i(v)\widehat{\xi}(Z_{1i})$ with $\mathbf{c}(v) \coloneqq \widehat{\mathbf{\Psi}}_1(\widehat{\mathbf{\Psi}}_1^\top\widehat{\mathbf{\Psi}}_1 + n\lambda_1\mathbf{I}_M)^{-1}\psi_{\widehat{\theta}_1}(v)$
return $\widehat{\mu}_{Y^1|V}$

Below, we derive the explicit formulae for the density estimators based on all three proposed methods. For the formulae to be explicit, we denote the set of the treated units in $\mathcal{D}_0$ by $\mathcal{D}_0^{A=1} \coloneqq \{Z_{0(1)}, \ldots, Z_{0(m)}\} = \{Z_{0i} \in \mathcal{D}_0 \mid A_{0i} = 1\}$. We denote $\mathbf{\Phi}_0 \coloneqq (\phi(Y_{0(1)}), \ldots, \phi(Y_{0(m)}))^\top$. Define $\mathbf{k}_{Y_1}(y) \coloneqq (k_{\mathcal{Y}}(y, Y_{11}), \ldots, k_{\mathcal{Y}}(y, Y_{1n}))^\top$ and $\mathbf{k}_{Y_0}(y) \coloneqq (k_{\mathcal{Y}}(y, Y_{0(1)}), \ldots, k_{\mathcal{Y}}(y, Y_{0(m)}))^\top$. For brevity, we also denote $\omega_i \coloneqq A_{1i}/\widehat{\pi}(X_{1i})$, $\mathbf{D}_\omega \coloneqq \mathrm{diag}(\omega_1, \ldots, \omega_n)$, and $\mathbf{D}_{1-\omega} \coloneqq \mathbf{I}_n - \mathbf{D}_\omega$.

C.1 Inference with Ridge Regression Estimator

We recall from (5) the Ridge Regression estimator $\widehat{\mu}_{\mathrm{RR}}(v) = \widehat{\mathbf{\Xi}}(\mathbf{K}_V + n\lambda_1\mathbf{I}_n)^{-1}\mathbf{k}_V(v)$. We assume the first-stage CME estimator is also the kernel ridge regression fitted on the treatment group in $\mathcal{D}_0$, denoted by $\widehat{\mu}_{0,\mathrm{RR}}(x) \coloneqq \mathbf{\Phi}_0(\mathbf{K}_{X_0} + m\lambda_0\mathbf{I}_m)^{-1}\mathbf{k}_{X_0}(x)$ where $\mathbf{K}_{X_0} = (k_{\mathcal{X}}(X_{0(i)}, X_{0(j)}))_{i,j=1}^m$ and $\mathbf{k}_{X_0}(x) = (k_{\mathcal{X}}(x, X_{0(1)}), \ldots, k_{\mathcal{X}}(x, X_{0(m)}))^\top$. Constructing the pseudo-outcome $\widehat{\xi}(Z_{1i}) = \omega_i(\phi(Y_{1i}) - \widehat{\mu}_{0,\mathrm{RR}}(X_{1i})) + \widehat{\mu}_{0,\mathrm{RR}}(X_{1i})$, we obtain the CCD estimator:

$$\begin{aligned}
\widehat{\mathfrak{p}}_{\mathrm{RR}}^1(y|v) &= \Big\langle \widehat{\mathbf{\Xi}}(\mathbf{K}_V + n\lambda_1\mathbf{I}_n)^{-1}\mathbf{k}_V(v), \phi(y)\Big\rangle_{\mathcal{H}_\mathcal{Y}} \\
&= (\langle\widehat{\xi}(Z_{11}), \phi(y)\rangle_{\mathcal{H}_\mathcal{Y}}, \ldots, \langle\widehat{\xi}(Z_{1n}), \phi(y)\rangle_{\mathcal{H}_\mathcal{Y}})^\top(\mathbf{K}_V + n\lambda_1\mathbf{I}_n)^{-1}\mathbf{k}_V(v) \\
&= \big[\mathbf{k}_{Y_1}(y)^\top\mathbf{D}_\omega + \mathbf{k}_{Y_0}(y)^\top(\mathbf{K}_{X_0} + m\lambda_0\mathbf{I}_m)^{-1}\mathbf{K}_{X_0X_1}\mathbf{D}_{1-\omega}\big](\mathbf{K}_V + n\lambda_1\mathbf{I}_n)^{-1}\mathbf{k}_V(v),
\end{aligned} \tag{13}$$

where $\mathbf{K}_{X_0X_1} \coloneqq (k_{\mathcal{X}}(X_{0(i)}, X_{1j}))_{1\le i\le m, 1\le j\le n}$ is a cross-kernel matrix.

Algorithm 3 Neural-Kernel Estimator

Require: Dataset $\mathcal{D} = \{Z_i\}_{i=1}^{2n}$ where $Z_i = (X_i, A_i, Y_i)$
Require: Kernel $k_{\mathcal{Y}}$ on $\mathcal{Y}$
Require: Propensity score estimation algorithm Alg_π
Require: Neural network parameter spaces Θ_0, Θ_1
Require: Grid points $\widetilde{\mathcal{Y}}_M = \{\widetilde{y}_1, \ldots, \widetilde{y}_M\} \subset \mathcal{Y}$

Randomly split $\mathcal{D}$ into two disjoint sets: $\mathcal{D}_0$ and $\mathcal{D}_1$, each of size n
$\{Z_{0(i)}\}_{i=1}^m \leftarrow \{Z_{0i} \in \mathcal{D}_0 \mid A_{0i} = 1\}$
Stage 1: Nuisance Estimation on $\mathcal{D}_0$
  $\widehat{\pi} \leftarrow \mathsf{Alg}_\pi(\mathcal{D}_0)$
  Compute grid Gram matrix $\mathbf{K}_M \in \mathbb{R}^{M\times M}$ with $(\mathbf{K}_M)_{jl} = k_{\mathcal{Y}}(\widetilde{y}_j, \widetilde{y}_l)$
  Initialize neural network $g_\theta : \mathcal{X} \to \mathbb{R}^M$ with $\theta \in \Theta_0$
  $\mathcal{L}_0(\theta) \leftarrow \frac{1}{m}\sum_{i=1}^m \Big[g_\theta(X_{0(i)})^\top \mathbf{K}_M g_\theta(X_{0(i)}) - 2\sum_{j=1}^M g_\theta(X_{0(i)})_j k_{\mathcal{Y}}(\widetilde{y}_j, Y_{0(i)})\Big]$
  $\widehat{\theta}_0 \leftarrow \arg\min_{\theta\in\Theta_0} \mathcal{L}_0(\theta)$
Stage 2: Pseudo-Outcome Regression on $\mathcal{D}_1$
  Initialize neural network $f_\theta : \mathcal{V} \to \mathbb{R}^M$ with $\theta \in \Theta_1$
  for $i = 1$ to n **do**
    $\mathbf{k}_i \leftarrow (k_{\mathcal{Y}}(\widetilde{y}_1, Y_{1i}), \ldots, k_{\mathcal{Y}}(\widetilde{y}_M, Y_{1i}))^\top$
    $\mathbf{b}_i \leftarrow \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\mathbf{k}_i + \Big(1 - \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\Big)\mathbf{K}_M g_{\widehat{\theta}_0}(X_{1i})$
    $\mathcal{L}_{1,i}(\theta) \leftarrow f_\theta(V_{1i})^\top \mathbf{K}_M f_\theta(V_{1i}) - 2 f_\theta(V_{1i})^\top \mathbf{b}_i$
  end for
  $\mathcal{L}_1(\theta) \leftarrow \frac{1}{n}\sum_{i=1}^n \mathcal{L}_{1,i}(\theta)$
  $\widehat{\theta}_1 \leftarrow \arg\min_{\theta\in\Theta_1} \mathcal{L}_1(\theta)$
  $\widehat{\mu}_{Y^1|V}(v) \leftarrow \sum_{j=1}^M f_{\widehat{\theta}_1}(v)_j \phi(\widetilde{y}_j)$
return $\widehat{\mu}_{Y^1|V}$

C.2 Inference with Deep Feature Estimator

Recall from (7) the Deep Feature estimator $\widehat{\mu}_{\mathrm{DF}}(v) = \widehat{\boldsymbol{\Xi}}\boldsymbol{\Psi}_{\widehat{\theta}}(\boldsymbol{\Psi}_{\widehat{\theta}}^\top \boldsymbol{\Psi}_{\widehat{\theta}} + n\lambda_1 \mathbf{I}_M)^{-1}\psi_{\widehat{\theta}}(v)$. Analogously, with a learned feature map $\psi_{0,\widehat{\theta}_0}$, denoting $\widehat{\boldsymbol{\Psi}}_0 \coloneqq (\psi_{0,\widehat{\theta}_0}(X_{0(1)}), \ldots, \psi_{0,\widehat{\theta}_0}(X_{0(m)}))^\top$, the first-stage Deep Feature estimator of the CME is given by:

$$\widehat{\mu}_{0,\mathrm{DF}}(x) \coloneqq \boldsymbol{\Phi}_0 \widehat{\boldsymbol{\Psi}}_0 (\widehat{\boldsymbol{\Psi}}_0^\top \widehat{\boldsymbol{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \psi_{0,\widehat{\theta}_0}(x).$$

Constructing the pseudo-outcome $\widehat{\xi}(Z_{1i})$ based on $\widehat{\mu}_{0,\mathrm{DF}}(X_{1i})$, we obtain the CCD estimator:

$$\begin{aligned}
&\widehat{\mathfrak{p}}^1_{\mathrm{DF}}(y|v) \\
&\quad = \Big\langle \widehat{\boldsymbol{\Xi}}\boldsymbol{\Psi}_{\widehat{\theta}}(\boldsymbol{\Psi}_{\widehat{\theta}}^\top \boldsymbol{\Psi}_{\widehat{\theta}} + n\lambda_1 \mathbf{I}_M)^{-1}\psi_{\widehat{\theta}}(v), \phi(y) \Big\rangle_{\mathcal{H}_{\mathcal{Y}}} \\
&\quad = \Big[\mathbf{k}_{Y_1}(y)^\top \mathbf{D}_\omega + \mathbf{k}_{Y_0}(y)^\top \widehat{\boldsymbol{\Psi}}_0 (\widehat{\boldsymbol{\Psi}}_0^\top \widehat{\boldsymbol{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \widehat{\boldsymbol{\Psi}}_{01} \mathbf{D}_{1-\omega}\Big] \\
&\qquad \boldsymbol{\Psi}_{\widehat{\theta}}(\boldsymbol{\Psi}_{\widehat{\theta}}^\top \boldsymbol{\Psi}_{\widehat{\theta}} + n\lambda_1 \mathbf{I}_M)^{-1}\psi_{\widehat{\theta}}(v),
\end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{\Psi}}_{01} \coloneqq (\psi_{0,\widehat{\theta}_0}(X_{11}), \ldots, \psi_{0,\widehat{\theta}_0}(X_{1n}))$ is the first-stage network evaluated on the second-stage covariates.

C.3 Inference with Neural-Kernel Estimator

Since the Neural-Kernel estimator $\widehat{\mu}_{\mathrm{NK}}(v) = \sum_{j=1}^M f_{\widehat{\theta}}(v)_j \phi(\widetilde{y}_j)$ is expressed as a linear combination of feature maps, the CCD estimator can be easily computed:

$$\widehat{\mathfrak{p}}^1_{\mathrm{NK}}(y|v) = \Big\langle \sum_{j=1}^M f_{\widehat{\theta}}(v)_j \phi(\widetilde{y}_j), \phi(y) \Big\rangle_{\mathcal{H}_{\mathcal{Y}}} = \sum_{j=1}^M f_{\widehat{\theta}}(v)_j k_{\mathcal{Y}}(\widetilde{y}_j, y). \tag{15}$$

This makes inference with the Neural-Kernel estimator significantly faster compared to the previous two estimators, since there is no need for calculation or inversion of large matrices.

D Baseline Estimators for Conditional Counterfactual Densities

In this section, we derive the CCD estimators for the two ablation baselines used in the experiments: the Inverse Propensity Weighting (IPW) estimator and the Plug-in (PI) estimator. These estimators use only one component of the doubly robust pseudo-outcome $\widehat{\xi}(Z)$ from (3). Recall that the doubly robust pseudo-outcome can be written as:

$$\widehat{\xi}(Z_{1i}) = \omega_i \phi(Y_{1i}) - \omega_i \widehat{\mu}_0(X_{1i}) + \widehat{\mu}_0(X_{1i}),$$

where $\omega_i = A_{1i}/\widehat{\pi}(X_{1i})$. The IPW and PI estimators correspond to retaining only the first or last term on the right-hand side, respectively. For each baseline, we derive the CCD estimators for all three methods by specializing the doubly robust density formulae from Appendix C.

D.1 Inverse Propensity Weighting Estimator

The Inverse Propensity Weighting (IPW) estimator relies solely on the propensity score to correct for the covariate shift between the observational and counterfactual distributions. The pseudo-outcome is:

$$\widehat{\xi}_{\mathrm{IPW}}(Z_{1i}) = \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\phi(Y_{1i}) = \omega_i \phi(Y_{1i}).$$

Since $\widehat{\xi}_{\mathrm{IPW}}$ does not involve the outcome model $\widehat{\mu}_0$, the inner product with $\phi(y)$ reduces to:

$$\langle \widehat{\xi}_{\mathrm{IPW}}(Z_{1i}), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} = \omega_i k_\mathcal{Y}(Y_{1i}, y), \tag{16}$$

yielding the vector $\mathbf{D}_\omega \mathbf{k}_{Y_1}(y)$ when collected over $i = 1, \ldots, n$.

The first-stage training remains the same for all methods. We describe the second stage for each method below.

Ridge Regression. Replacing the bracket term in (13) by $\mathbf{D}_\omega \mathbf{k}_{Y_1}(y)$, we obtain:

$$\widehat{\mathfrak{p}}^1_{\mathrm{IPW,RR}}(y|v) = \mathbf{k}_{Y_1}(y)^\top \mathbf{D}_\omega (\mathbf{K}_V + n\lambda_1 \mathbf{I}_n)^{-1} \mathbf{k}_V(v).$$

Deep Feature. To train the second-stage network ψ_θ, we substitute the IPW pseudo-outcome Gram matrix into the loss (8). The entries of this matrix are:

$$(\mathbf{K}^{\mathrm{IPW}}_{\widehat{\xi}})_{ij} = \langle \widehat{\xi}_{\mathrm{IPW}}(Z_{1i}), \widehat{\xi}_{\mathrm{IPW}}(Z_{1j}) \rangle_{\mathcal{H}_\mathcal{Y}} = \omega_i \omega_j \, k_\mathcal{Y}(Y_{1i}, Y_{1j}),$$

which in matrix form reads $\mathbf{K}^{\mathrm{IPW}}_{\widehat{\xi}} = \mathbf{D}_\omega \mathbf{K}_{Y_1} \mathbf{D}_\omega$. Let $\psi_{\widehat{\theta}}$ be the learned feature map. Replacing the bracket term in (14) by $\mathbf{D}_\omega \mathbf{k}_{Y_1}(y)$ gives:

$$\widehat{\mathfrak{p}}^1_{\mathrm{IPW,DF}}(y|v) = \mathbf{k}_{Y_1}(y)^\top \mathbf{D}_\omega \boldsymbol{\Psi}_{\widehat{\theta}} (\boldsymbol{\Psi}_{\widehat{\theta}}^\top \boldsymbol{\Psi}_{\widehat{\theta}} + n\lambda_1 \mathbf{I}_M)^{-1} \psi_{\widehat{\theta}}(v).$$

Neural-Kernel. The vector $\mathbf{b}_i \in \mathbb{R}^M$ for the Neural-Kernel loss (9) simplifies to:

$$(\mathbf{b}_i)_j = \left\langle \phi(\widetilde{y}_j), \widehat{\xi}_{\mathrm{IPW}}(Z_{1i}) \right\rangle_{\mathcal{H}_\mathcal{Y}} = \frac{A_{1i}}{\widehat{\pi}(X_{1i})} k_\mathcal{Y}(\widetilde{y}_j, Y_{1i}),$$

that is, $\mathbf{b}_i = \omega_i \mathbf{k}_i$. Substituting into (9) yields the training loss:

$$\widehat{\mathcal{L}}_{\mathrm{IPW,NK}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[f_\theta(V_{1i})^\top \mathbf{K}_M f_\theta(V_{1i}) - 2\omega_i \, f_\theta(V_{1i})^\top \mathbf{k}_i \right].$$

Let $\widehat{\theta}$ be the parameter obtained from training the model with this loss. The density estimate follows from (15): $\widehat{\mathfrak{p}}^1_{\mathrm{IPW,NK}}(y|v) = \sum_{j=1}^{M} f_{\widehat{\theta}}(v)_j k_\mathcal{Y}(\widetilde{y}_j, y)$.

D.2 Plug-in Density Estimator

The Plug-in (PI) estimator relies solely on the conditional mean embedding of the outcome, ignoring the propensity scores. The pseudo-outcome is the prediction of the first-stage model:

$$\widehat{\xi}_{\mathrm{PI}}(Z_{1i}) = \widehat{\mu}_0(X_{1i}).$$

The inner product with $\phi(y)$ is:

$$\langle \widehat{\xi}_{\mathrm{PI}}(Z_{1i}), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} = \langle \widehat{\mu}_0(X_{1i}), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} = \widehat{\mu}_0(X_{1i})(y). \tag{17}$$

The first-stage training remains the same for all methods. We describe the second stage for each method below.

Ridge Regression. Assume the first-stage estimator is the Ridge Regression estimator fitted on the treatment group, $\widehat{\mu}_{0,\mathrm{RR}}(x) = \mathbf{\Phi}_0(\mathbf{K}_{X_0} + m\lambda_0 \mathbf{I}_m)^{-1}\mathbf{k}_{X_0}(x)$. We then calculate (17):

$$\langle \widehat{\mu}_{0,\mathrm{RR}}(X_{1i}), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbf{k}_{Y_0}(y)^\top (\mathbf{K}_{X_0} + m\lambda_0 \mathbf{I}_m)^{-1} \mathbf{k}_{X_0}(X_{1i}).$$

Collecting over $i = 1, \ldots, n$ yields the vector $\mathbf{K}_{X_0 X_1}^\top (\mathbf{K}_{X_0} + m\lambda_0 \mathbf{I}_m)^{-1} \mathbf{k}_{Y_0}(y)$. Replacing the bracket term in (13) by this vector, we obtain:

$$\widehat{\mathfrak{p}}^1_{\mathrm{PI,RR}}(y|v) = \mathbf{k}_{Y_0}(y)^\top (\mathbf{K}_{X_0} + m\lambda_0 \mathbf{I}_m)^{-1} \mathbf{K}_{X_0 X_1} (\mathbf{K}_V + n\lambda_1 \mathbf{I}_n)^{-1} \mathbf{k}_V(v).$$

Deep Feature. Assume the first-stage estimator is the Deep Feature estimator with the learned feature map $\psi_{0,\widehat{\theta}_0}$. To train the second-stage network, we substitute the Gram matrix $\mathbf{K}^{\mathrm{PI}}_{\widehat{\xi}} := (\langle \widehat{\mu}_0(X_{1i}), \widehat{\mu}_0(X_{1j}) \rangle_{\mathcal{H}_\mathcal{Y}})_{i,j=1}^n$ into the loss (8). Since $\widehat{\mu}_0(X_{1i}) = \mathbf{\Phi}_0 \widehat{\mathbf{\Psi}}_0 (\widehat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \psi_{0,\widehat{\theta}_0}(X_{1i})$, we have:

$$\mathbf{K}^{\mathrm{PI}}_{\widehat{\xi}} = \widehat{\mathbf{\Psi}}_{01}^\top (\widehat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \widehat{\mathbf{\Psi}}_0^\top \mathbf{K}_{Y_0} \widehat{\mathbf{\Psi}}_0 (\widehat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \widehat{\mathbf{\Psi}}_{01}.$$

After the second-stage optimization, we calculate (17):

$$\langle \widehat{\mu}_{0,\mathrm{DF}}(X_{1i}), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbf{k}_{Y_0}(y)^\top \widehat{\mathbf{\Psi}}_0 (\widehat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \psi_{0,\widehat{\theta}_0}(X_{1i}).$$

Collecting over $i = 1, \ldots, n$ and replacing the bracket term in (14) by this vector, we obtain:

$$\widehat{\mathfrak{p}}^1_{\mathrm{PI,DF}}(y|v) = \mathbf{k}_{Y_0}(y)^\top \widehat{\mathbf{\Psi}}_0 (\widehat{\mathbf{\Psi}}_0^\top \widehat{\mathbf{\Psi}}_0 + m\lambda_0 \mathbf{I}_M)^{-1} \widehat{\mathbf{\Psi}}_{01} \mathbf{\Psi}_{\widehat{\theta}} (\mathbf{\Psi}_{\widehat{\theta}}^\top \mathbf{\Psi}_{\widehat{\theta}} + n\lambda_1 \mathbf{I}_M)^{-1} \psi_{\widehat{\theta}}(v).$$

Neural-Kernel. Assume the first-stage estimator is a Neural-Kernel estimator $\widehat{\mu}_{0,\mathrm{NK}}(x) := \sum_{j=1}^M g_{\widehat{\theta}_0}(x)_j \phi(\widetilde{y}_j)$ with the same grid points. The target vector $\mathbf{b}_i \in \mathbb{R}^M$ for the loss (9) simplifies to:

$$(\mathbf{b}_i)_j = \langle \phi(\widetilde{y}_j), \widehat{\mu}_{0,\mathrm{NK}}(X_{1i}) \rangle_{\mathcal{H}_\mathcal{Y}} = \sum_{l=1}^M g_{\widehat{\theta}_0}(X_{1i})_l \, k_\mathcal{Y}(\widetilde{y}_l, \widetilde{y}_j),$$

that is, $\mathbf{b}_i = \mathbf{K}_M g_{\widehat{\theta}_0}(X_{1i})$. Substituting into (9) yields the training loss:

$$\widehat{\mathcal{L}}_{\mathrm{PI,NK}}(\theta) = \frac{1}{n} \sum_{i=1}^n \Big[f_\theta(V_{1i})^\top \mathbf{K}_M f_\theta(V_{1i}) - 2 f_\theta(V_{1i})^\top \mathbf{K}_M g_{\widehat{\theta}_0}(X_{1i}) \Big].$$

Let $\widehat{\theta}$ be the parameter obtained from training the model with this loss. The density estimate follows from (15): $\widehat{\mathfrak{p}}^1_{\mathrm{PI,NK}}(y|v) = \sum_{j=1}^M f_{\widehat{\theta}}(v)_j k_\mathcal{Y}(\widetilde{y}_j, y)$.

E Applications of Theorem 1

We provide three applications of Theorem 1 to three particular tasks, namely density estimation, hypothesis testing, and sampling of conditional counterfactual distributions. For each task, we demonstrate an application of our meta-estimator bound. Specifically, we bound the error of each task in terms of $\mathbb{E}\Big[\|\widehat{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}} \Big]$, which allows us to apply Theorem 1 to bound this error by the first-stage and second-stage errors.

E.1 Density Estimation

Let $k_\mathcal{Y}(y, y') = \sigma^{-d_y} \rho((y - y')/\sigma)$ with a bandwidth $\sigma > 0$, where $\rho \geq 0$ is a positive definite function on $\mathbb{R}^{d_y}$ with $\int \rho(u)\, \mathrm{d}u = 1$ and $\int \|u\|^2 \rho(u)\, \mathrm{d}u < \infty$. As in Appendix C, we can use the density estimator $\widehat{\mathfrak{p}}^1(y|v) = \langle \widehat{\mu}_{Y^1|V}(v), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}}$ to estimate the conditional counterfactual density $\mathfrak{p}^1(\cdot|v)$. The following corollary provides a bound on the error of $\widehat{\mathfrak{p}}^1(\cdot|v)$ as an estimator of $\mathfrak{p}^1(\cdot|v)$.

Corollary 1 (Density Estimation)**.** *Let $k_{\mathcal{Y}}$ be as above. Assume that, for every $v \in \mathcal{V}$, the function $y \mapsto \mathfrak{p}^1(y|v)$, extended by zero outside $\mathcal{Y}$, is twice differentiable on $\mathbb{R}^{d_y}$ with $\sup_y \|\nabla_y^2 \mathfrak{p}^1(y|v)\|_{\mathrm{op}} \le L_{\mathfrak{p}}$ for some $L_{\mathfrak{p}} > 0$, where $\|\cdot\|_{\mathrm{op}}$ is the matrix operator norm. Then,*

$$\mathbb{E}\left[\sup_{y\in\mathcal{Y}} \left|\widehat{\mathfrak{p}}^1(y|V) - \mathfrak{p}^1(y|V)\right|^2\right] \lesssim \frac{1}{\sigma^{d_y}} \mathbb{E}\big[\|\widehat{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_{\mathcal{Y}}}^2\big] + \sigma^4.$$

Proof. By the reproducing property and the symmetry $\rho(-u) = \rho(u)$, substituting $y' = y - \sigma u$ yields:

$$\bar{\mathfrak{p}}^1(y|v) \coloneqq \langle \mu_{Y^1|V}(v), \phi(y)\rangle_{\mathcal{H}_{\mathcal{Y}}} = \frac{1}{\sigma^{d_y}} \int \rho\left(\frac{y - y'}{\sigma}\right) \mathfrak{p}^1(y'|v)\, \mathrm{d}y' = \int \rho(u) \mathfrak{p}^1(y - \sigma u|v)\, \mathrm{d}u.$$

We then decompose the target error as $\widehat{\mathfrak{p}}^1 - \mathfrak{p}^1 = (\widehat{\mathfrak{p}}^1 - \bar{\mathfrak{p}}^1) + (\bar{\mathfrak{p}}^1 - \mathfrak{p}^1)$ and apply the triangle inequality. By the reproducing property and the Cauchy-Schwarz inequality, we have:

$$\begin{aligned}
\left|\widehat{\mathfrak{p}}^1(y|v) - \bar{\mathfrak{p}}^1(y|v)\right| &= \left|\langle \widehat{\mu}_{Y^1|V}(v) - \mu_{Y^1|V}(v), \phi(y)\rangle_{\mathcal{H}_{\mathcal{Y}}}\right| \\
&\le \sqrt{k_{\mathcal{Y}}(y,y)}\, \|\widehat{\mu}_{Y^1|V}(v) - \mu_{Y^1|V}(v)\|_{\mathcal{H}_{\mathcal{Y}}} \\
&= \sqrt{\frac{\rho(0)}{\sigma^{d_y}}}\, \|\widehat{\mu}_{Y^1|V}(v) - \mu_{Y^1|V}(v)\|_{\mathcal{H}_{\mathcal{Y}}}. \qquad (18)
\end{aligned}$$

Next, we bound $\bar{\mathfrak{p}}^1 - \mathfrak{p}^1$. Since $\int \rho(u)\,\mathrm{d}u = 1$ and $\int u\,\rho(u)\,\mathrm{d}u = 0$ by the symmetry of ρ, a second-order Taylor expansion of $\mathfrak{p}^1(y - \sigma u|v)$ around y yields:

$$\left|\bar{\mathfrak{p}}^1(y|v) - \mathfrak{p}^1(y|v)\right| = \left|\int \rho(u)\big(\mathfrak{p}^1(y - \sigma u|v) - \mathfrak{p}^1(y|v)\big)\,\mathrm{d}u\right| \le \frac{L_{\mathfrak{p}}\sigma^2}{2} \int \|u\|^2 \rho(u)\,\mathrm{d}u. \qquad (19)$$

The desired bound then immediately follows from (18) and (19). □

E.2 Testing

We denote by $\mu_{Y^0|V}$ the CCME of the counterfactual Y^0: $\mu_{Y^0|V}(v) \coloneqq \mathbb{E}[\phi(Y^0)|V = v]$. The meta-estimator $\widehat{\mu}_{Y^0|V}$ of $\mu_{Y^0|V}$ is $\widehat{\mu}_{Y^1|V}$ with A and $\widehat{\pi}$ replaced by $1 - A$ and $1 - \widehat{\pi}$, and with $\widehat{\mu}_0$ replaced by a first-stage estimator $\widehat{\nu}_0$ of $\nu_0(x) \coloneqq \mathbb{E}[\phi(Y)|X = x, A = 0]$. We randomly split the indices of $\mathcal{D}_1$ into two sets $\mathcal{I}_1$ and $\mathcal{I}_2$ of size $n/2$. Let $\widehat{\mu}^{(1)}_{Y^1|V}$ and $\widehat{\mu}^{(1)}_{Y^0|V}$ be the meta-estimators with the second stage computed on $\mathcal{I}_1$, and $\widehat{\mu}^{(2)}_{Y^1|V}$ and $\widehat{\mu}^{(2)}_{Y^0|V}$ be those on $\mathcal{I}_2$. We then define:

$$\widehat{\Delta}^{(1)}(v) \coloneqq \widehat{\mu}^{(1)}_{Y^1|V}(v) - \widehat{\mu}^{(1)}_{Y^0|V}(v), \qquad \widehat{\Delta}^{(2)}(v) \coloneqq \widehat{\mu}^{(2)}_{Y^1|V}(v) - \widehat{\mu}^{(2)}_{Y^0|V}(v).$$

To test the null hypothesis H_0 that $P(Y^1|V = v) = P(Y^0|V = v)$ for almost every v, we use the cross U-statistic of [KR24], which [MTRK23] apply to the marginal distributions of Y^1 and Y^0:

$$\widehat{T}(v) \coloneqq \langle \widehat{\Delta}^{(1)}(v), \widehat{\Delta}^{(2)}(v)\rangle_{\mathcal{H}_{\mathcal{Y}}}. \qquad (20)$$

This statistic estimates the squared *maximum mean discrepancy* between $P(Y^1|V = v)$ and $P(Y^0|V = v)$:

$$T(v) \coloneqq \mathrm{MMD}^2(P(Y^1|V = v), P(Y^0|V = v)) = \|\mu_{Y^1|V}(v) - \mu_{Y^0|V}(v)\|_{\mathcal{H}_{\mathcal{Y}}}^2.$$

For a characteristic kernel, $T(v) = 0$ if and only if $P(Y^1|V = v) = P(Y^0|V = v)$. The following corollary provides a bound on the estimation error of the statistic.

Corollary 2 (Testing)**.** *The statistic $\widehat{T}$ satisfies:*

$$\begin{aligned}
\mathbb{E}\left[\left|\widehat{T}(V) - T(V)\right|\right] \lesssim\ & \mathbb{E}\left[\|\widehat{\mu}^{(1)}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] + \mathbb{E}\left[\|\widehat{\mu}^{(1)}_{Y^0|V}(V) - \mu_{Y^0|V}(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&+ \mathbb{E}[T(V)]^{1/2} \Bigg(\mathbb{E}\left[\|\widehat{\mu}^{(1)}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] \\
&\qquad + \mathbb{E}\left[\|\widehat{\mu}^{(1)}_{Y^0|V}(V) - \mu_{Y^0|V}(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]\Bigg)^{1/2}.
\end{aligned}$$

Proof. This follows from the following application of the Cauchy–Schwarz inequality:

$$\begin{aligned}\left|\widehat{T}(v) - T(v)\right| \le{}& \|\widehat{\Delta}^{(1)}(v) - \mu_{Y^1|V}(v) + \mu_{Y^0|V}(v)\|_{\mathcal{H}_\mathcal{Y}} \|\widehat{\Delta}^{(2)}(v) - \mu_{Y^1|V}(v) + \mu_{Y^0|V}(v)\|_{\mathcal{H}_\mathcal{Y}} \\ &+ T(v)^{1/2} \|\widehat{\Delta}^{(1)}(v) - \mu_{Y^1|V}(v) + \mu_{Y^0|V}(v)\|_{\mathcal{H}_\mathcal{Y}} \\ &+ T(v)^{1/2} \|\widehat{\Delta}^{(2)}(v) - \mu_{Y^1|V}(v) + \mu_{Y^0|V}(v)\|_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

It remains to take expectations and apply the Cauchy–Schwarz and triangle inequalities, since $(V, \widehat{\Delta}^{(1)}(V))$ and $(V, \widehat{\Delta}^{(2)}(V))$ have the same distribution. □

E.3 Sampling

Suppose that $\mathcal{Y}$ is compact and that $k_\mathcal{Y}$ is continuous. Kernel herding [Wel09, CWS10] with the target $\widehat{\mu}_{Y^1|V}(v)$ generates points $y_1, y_2, \ldots \in \mathcal{Y}$ by

$$y_{t+1} \in \arg\max_{y \in \mathcal{Y}} \left[\langle \widehat{\mu}_{Y^1|V}(v), \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} - \frac{1}{t+1} \sum_{s=1}^{t} k_\mathcal{Y}(y_s, y) \right], \qquad t = 0, 1, \ldots. \tag{21}$$

When the target is the embedding of a distribution and lies in the relative interior of the convex hull $\mathcal{M} \coloneqq \operatorname{conv}\{\phi(y) : y \in \mathcal{Y}\}$, [CWS10, Proposition 1] shows that the error $\|\widehat{\mu}_{Y^1|V}(v) - \frac{1}{N}\sum_{t=1}^{N} \phi(y_t)\|_{\mathcal{H}_\mathcal{Y}}$ decays as $O(N^{-1})$. However, when $\mathcal{H}_\mathcal{Y}$ is infinite-dimensional, as for the Gaussian kernel, the relative interior condition fails [BLO12, Proposition 2]. In fact, the target $\widehat{\mu}_{Y^1|V}(v)$ might not even lie in $\mathcal{M}$; an example is our own CCME estimator, since the pseudo-outcome (3) puts the negative weight $1 - 1/\widehat{\pi}(x)$ on $\widehat{\mu}_0(x)$ when $a = 1$. Here, we demonstrate that our meta-estimator bound makes it possible to bound the error of the herding points with respect to $P(Y^1|V = v)$ in this regime, albeit with a slower rate of $O(N^{-1/2})$.

Corollary 3 (Sampling). *Let $\mathcal{Y}$ be compact, and let $k_\mathcal{Y}$ be continuous with $\sup_{y \in \mathcal{Y}} k_\mathcal{Y}(y, y) \le B_k$. Then, for every $v \in \mathcal{V}$ and every $N \ge 1$, the empirical distribution $\widehat{P}_N$ of the points $y_1, \ldots, y_N$ generated by (21) satisfies:*

$$\operatorname{MMD}\Big(\widehat{P}_N, P(Y^1|V = v)\Big) \lesssim \|\widehat{\mu}_{Y^1|V}(v) - \mu_{Y^1|V}(v)\|_{\mathcal{H}_\mathcal{Y}} + \frac{\sqrt{B_k} + \|\widehat{\mu}_{Y^1|V}(v)\|_{\mathcal{H}_\mathcal{Y}}}{\sqrt{N}}.$$

Proof. We fix v and denote $\widehat{\mu} \coloneqq \widehat{\mu}_{Y^1|V}(v)$, $\mu \coloneqq \mu_{Y^1|V}(v)$, $\Delta \coloneqq \|\widehat{\mu} - \mu\|_{\mathcal{H}_\mathcal{Y}}$, and $R \coloneqq \sqrt{B_k} + \|\widehat{\mu}\|_{\mathcal{H}_\mathcal{Y}}$. Since $\frac{1}{N}\sum_{t=1}^{N} \phi(y_t)$ is the embedding of $\widehat{P}_N$, the triangle inequality yields:

$$\operatorname{MMD}\big(\widehat{P}_N, P(Y^1|V = v)\big) \le \Delta + \left\| \widehat{\mu} - \frac{1}{N} \sum_{t=1}^{N} \phi(y_t) \right\|_{\mathcal{H}_\mathcal{Y}}. \tag{22}$$

To bound the second term on the right-hand side, we follow the idea from [CWS10]. First, we define the weight vectors:

$$w_t \coloneqq (t+1)\widehat{\mu} - \sum_{s=1}^{t} \phi(y_s), \qquad t \ge 0.$$

Since $w_N - w_0 = N\widehat{\mu} - \sum_{t=1}^{N} \phi(y_t)$, the second term on the right-hand side of (22) is at most $(\|w_N\|_{\mathcal{H}_\mathcal{Y}} + \|w_0\|_{\mathcal{H}_\mathcal{Y}})/N$. Since $w_0 = \widehat{\mu}$, it suffices to bound $\|w_N\|_{\mathcal{H}_\mathcal{Y}}$, which we do by induction on t through a recursion that relates $\|w_{t+1}\|_{\mathcal{H}_\mathcal{Y}}$ to $\|w_t\|_{\mathcal{H}_\mathcal{Y}}$.

Since the objective in (21) is $\langle w_t, \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}}/(t+1)$, we have $y_{t+1} \in \arg\max_{y \in \mathcal{Y}} \langle w_t, \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}}$. Denote $c_t \coloneqq \phi(y_{t+1}) - \widehat{\mu}$, which satisfies $w_{t+1} = w_t - c_t$ and $\|c_t\|_{\mathcal{H}_\mathcal{Y}} \le R$. Since y_{t+1} maximizes $\langle w_t, \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}}$ over $\mathcal{Y}$ and $P(Y^1|V = v)$ is supported in $\mathcal{Y}$, we have:

$$\begin{aligned}\langle w_t, c_t \rangle_{\mathcal{H}_\mathcal{Y}} &= \max_{y \in \mathcal{Y}} \langle w_t, \phi(y) \rangle_{\mathcal{H}_\mathcal{Y}} - \langle w_t, \widehat{\mu} \rangle_{\mathcal{H}_\mathcal{Y}} \\ &\ge \mathbb{E}_{Y \sim P(Y^1|V=v)}[\langle w_t, \phi(Y) \rangle_{\mathcal{H}_\mathcal{Y}}] - \langle w_t, \widehat{\mu} \rangle_{\mathcal{H}_\mathcal{Y}} \\ &= \langle w_t, \mu - \widehat{\mu} \rangle_{\mathcal{H}_\mathcal{Y}} \ge -\Delta \|w_t\|_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

Consequently, the norms of the weight vectors satisfy the recursion:

$$\|w_{t+1}\|_{\mathcal{H}_\mathcal{Y}}^2 = \|w_t\|_{\mathcal{H}_\mathcal{Y}}^2 - 2\langle w_t, c_t\rangle_{\mathcal{H}_\mathcal{Y}} + \|c_t\|_{\mathcal{H}_\mathcal{Y}}^2 \leq (\|w_t\|_{\mathcal{H}_\mathcal{Y}} + \Delta)^2 + R^2.$$

We now show by induction that $\|w_t\|_{\mathcal{H}_\mathcal{Y}} \leq \|w_0\|_{\mathcal{H}_\mathcal{Y}} + t\Delta + \sqrt{t}R$ for every $t \geq 0$. The case $t = 0$ is trivial. If the bound holds at t, then inserting this bound into the right-hand side of the last display yields:

$$\|w_{t+1}\|_{\mathcal{H}_\mathcal{Y}} \leq \left(\left(\|w_0\|_{\mathcal{H}_\mathcal{Y}} + (t+1)\Delta + \sqrt{t}R\right)^2 + R^2\right)^{1/2} \leq \|w_0\|_{\mathcal{H}_\mathcal{Y}} + (t+1)\Delta + \sqrt{t+1}R.$$

We thus complete the induction step. In particular, with $t = N$, we obtain:

$$\left\|\widehat{\mu} - \frac{1}{N}\sum_{t=1}^{N}\phi(y_t)\right\|_{\mathcal{H}_\mathcal{Y}} \leq \frac{\|w_N\|_{\mathcal{H}_\mathcal{Y}} + \|w_0\|_{\mathcal{H}_\mathcal{Y}}}{N} \leq \Delta + \frac{2\|\widehat{\mu}\|_{\mathcal{H}_\mathcal{Y}}}{N} + \frac{R}{\sqrt{N}} \leq \Delta + \frac{3R}{\sqrt{N}}.$$

Inserting this bound into the right-hand side of (22) yields the claim. □

F Experimental Details

We implement all matrix operations and automatic differentiation in `JAX` [BFH+18], all neural networks in `Equinox` [KG21], and all propensity score models in `scikit-learn` [PVG+11].

F.1 Computational Resources

All experiments were conducted on an internal workstation equipped with an Intel Core i9-9900K CPU (3.60 GHz, 8 physical cores), 32 GB of RAM, and a single NVIDIA GeForce RTX 2080 Ti GPU (11 GB VRAM).

F.2 Synthetic Data (Section 7.1)

We generate covariates $X \sim \mathcal{N}(\mathbf{1}_{10}, \mathbf{I}_{10}) \in \mathbb{R}^{10}$ with the conditioning variable $V = (X_1, X_2, X_3, X_4, X_5)$. The treatment assignment follows the propensity score: $P(A = 1 \,|\, X = x) = \pi(x) = 0.1 + 0.8 \cdot \mathbb{I}(x_1 \in [0, 2] \text{ and } x_6 \geq 1.5)$. Then we generate Y according to the equation:

$$\begin{aligned} Y &= 1 + X^\top\beta + A \cdot (2 + X^\top\gamma) + \epsilon, \\ \beta &\coloneqq (1, -0.5, 0.8, -0.7, 0.6, 1, 0.3, -0.2, 0.1, -0.3)^\top, \\ \gamma &\coloneqq (0.8, 0, 0, 0.6, 0, 2.0, 0.4, 0, 0, 0.2)^\top. \end{aligned}$$

Here, ϵ is a heteroscedastic bimodal noise with the following generation process:

$$\begin{aligned} \epsilon &= S + \epsilon', \\ S/15.0 &\sim \text{Bernoulli}(\text{sigmoid}(0.5X_1)), \\ \epsilon' &\sim \mathcal{N}\big(0, 0.25(1 + 0.5|X_1| + 0.3|X_5|)^2\big). \end{aligned}$$

The ground truth density $\mathfrak{p}^1(y|v)$ is computed analytically: since the covariates are independent, conditioning on $V = X_{1:5}$ leaves $X_{6:10} \sim \mathcal{N}(\mathbf{1}_5, \mathbf{I}_5)$ unchanged, so the conditional moments of Y^1 given $V = v$ are available in closed form.

For the qualitative assessment, we visualize 30 repeated runs of estimated counterfactual densities at $n = 10000$, evaluated at two covariate profiles: $v_1 \coloneqq (2.2, -0.2, 2.2, -0.2, 2.2)^\top$ and $v_2 \coloneqq (-0.2, 2.2, -0.2, 2.2, -0.2)^\top$. These induce different mixture probabilities $\text{sigmoid}(0.5X_1)$, which is 0.75 for v_1 and 0.48 for v_2, yielding distinct bimodal shapes.

Hyperparameters. For the Deep Feature and Neural-Kernel estimators, the learning rates scale linearly with sample size: $\kappa \times n/100$, where $\kappa = 2 \times 10^{-4}$ (Deep Feature) and $\kappa = 4 \times 10^{-4}$ (Neural-Kernel). Each random forest consists of 100 trees and has a maximum depth of 4. The full configurations are given in Table 1.

Table 1: Hyperparameter configurations for Section 7.1.

Hyperparameter	Ridge Regression	Deep Feature	Neural-Kernel
Model Parameters			
Kernel Bandwidth (σ)	2.0	2.0	2.0
Regularization ($n\lambda_1 = m\lambda_0$)	20.0	20.0	N/A
Grid/Feature Dimension (M)	N/A	20	20
Network Architecture	N/A	2 layers $\times$ 20 units, ReLU	
Training (SGD Optimizer)			
Learning Rate (Stage 1)	N/A	$2 \times 10^{-4} \times n/100$	$4 \times 10^{-4} \times n/100$
Learning Rate (Stage 2)	N/A	$2 \times 10^{-4} \times n/100$	$4 \times 10^{-4} \times n/100$
Momentum	N/A	0.9	0.9
Epochs (Stage 1)	N/A	6000	16000
Epochs (Stage 2)	N/A	1000	500
Batch Size	Full Batch	Full Batch	Full Batch

Evaluation. We generate 10000 test points $X_{\text{test}} \sim \mathcal{N}(\mathbf{1}_{10}, \mathbf{I}_{10})$ and evaluate $\widehat{\mathfrak{p}}^1(y|v_{\text{test}})$ at $v_{\text{test}} := \eta(X_{\text{test}})$ on 1000 grid points spanning $[\min(Y) - 2, \max(Y) + 2]$.

F.3 MNIST Experiment (Section 7.2)

Images are normalized to $[0, 1]$. Propensity scores are clipped to $[0.01, 0.99]$. The propensity score is estimated by logistic regression. We also use early stopping by splitting the data into 80% train and 20% validation sets, with patience set to 10 epochs. The hyperparameters are given in Table 2.

Table 2: Hyperparameter configuration for the MNIST experiment (Section 7.2).

Hyperparameter	Deep Feature Estimator
Feature Dimension (M)	1000
Network Architecture	2 layers $\times$ 100 units, ReLU
Kernel Bandwidth (σ)	1.0
Regularization ($n\lambda_1 = m\lambda_0$)	1.0
Training (SGD Optimizer)	
Learning Rate (Stage 1)	0.1
Learning Rate (Stage 2)	0.01
Momentum	0.9
Epochs (Stage 1)	100
Epochs (Stage 2)	100
Batch Size	32
Patience for Early Stopping	10

G Downstream Causal Tasks

We apply the doubly robust CCME estimator to three downstream tasks: (i) conditional value-at-risk (CVaR) treatment effects, (ii) a conditional two-sample test, and (iii) risk-sensitive policy learning.

Data generating process. In Tasks 1 and 3, we generate the data from the model

$$V \sim \text{Unif}[-2, 2], \quad U \sim \mathcal{N}(0, 1), \quad X = (V, U), \quad A \mid X \sim \text{Bernoulli}(\pi(X)),$$

$$Y^0 = V + 0.8U + \epsilon_0, \qquad Y^1 = V + 1.5 + S\,\delta(V) + 0.8U + \epsilon_1, \qquad \delta(v) := 2.2 + 0.9v,$$

where $\pi(X) = \mathsf{sigmoid}(0.5V + 1.1U)$ is truncated to $[0.05, 0.95]$, $S \sim \text{Unif}\{-1, +1\}$, $\epsilon_0, \epsilon_1 \sim \mathcal{N}(0, 0.7^2)$, all independent, and $Y = AY^1 + (1 - A)Y^0$. The confounder U affects both A and the potential outcomes but is not a component of V. Consequently, the One-Step estimator, which

adjusts for V only, is biased. Integrating out U, we obtain $P(Y^0|V=v) = \mathcal{N}(v, 0.8^2 + 0.7^2)$. In contrast, $P(Y^1|V=v)$ is the Gaussian mixture $\frac{1}{2}\mathcal{N}(v + 1.5 - \delta(v), 0.8^2 + 0.7^2) + \frac{1}{2}\mathcal{N}(v + 1.5 + \delta(v), 0.8^2 + 0.7^2)$. Thus, $\mathbb{E}[Y^1 - Y^0|V=v] = 1.5$ for every v, while the difference between the two conditional distributions varies with v through $\delta(v)$.

Estimators. We implement the four estimators as follows.

- CCME-DR: our proposed two-stage Ridge Regression estimator with a Gaussian kernel on $\mathcal{X}$, $\mathcal{V}$, and $\mathcal{Y}$, $n\lambda_1 = m\lambda_0 = 4$, a logistic regression estimate of the propensity score truncated to $[0.05, 0.95]$, and a split of the sample into two subsets.
- One-Step (CME/CoDiTE [PSSM21]): the single-stage kernel ridge regression of $\phi(Y)$ on V, fitted on each of the treated and control groups, with the same kernel and regularization. This estimator uses neither sample splitting nor a propensity score.
- CPME [ZBG25]: the augmented inverse propensity weighted embedding, with the same kernel, logistic regression, and regularization, and no sample split. This estimator fits the propensity score and the outcome model as in Stage 1 of CCME-DR. However, in contrast to the regression on V in Stage 2, this estimator takes the average of the pseudo-outcomes (3) over the sample.
- CATE-DR: the linear regression on V of the difference of the pseudo-outcomes (3) for Y^1 and Y^0 with $\phi(y)$ replaced by y, which estimates the conditional average treatment effect (CATE) $\mathbb{E}[Y^1 - Y^0 \mid V = v]$, with linear regressions of Y on X in the first stage, the same logistic regression, and a split of the sample into two subsets.

We estimate $\mu_{Y^0|V}$ with A replaced by $1 - A$, as in Appendix E. In Tasks 1 and 3, we draw a sample of size 2000 in each of 100 replications and evaluate the estimators at 100 equally spaced points $v \in [-2, 2]$.

Bandwidths. In every replication of every task, we select the bandwidths by fitting on 80% of the sample and minimizing an error on the held-out 20%, specifically the mean integrated squared error of the density of Y given V for the bandwidths on $\mathcal{V}$ and $\mathcal{Y}$, and the error of the regression of $\phi(Y)$ on X for the bandwidth on $\mathcal{X}$. We use the whole sample for this selection in Tasks 1 and 3, and only $\mathcal{D}_0$ in Task 2. All the kernel methods use the selected bandwidths.

Task 1: CVaR treatment effects. For $a \in \{0, 1\}$, we denote by $F_{Y^a|v}$ the cumulative distribution function of $P(Y^a|V=v)$. For $\alpha \in (0, 1)$, the conditional CVaR

$$\mathrm{CVaR}_\alpha(Y^a|v) := \frac{1}{\alpha}\int_0^\alpha F^{-1}_{Y^a|v}(u)\,\mathrm{d}u$$

is the mean of the lower α-tail of $P(Y^a|V=v)$. The CVaR treatment effect is $\mathrm{CVaR}_\alpha(Y^1|v) - \mathrm{CVaR}_\alpha(Y^0|v)$. To estimate this effect, we evaluate the estimated density $\widehat{\mathfrak{p}}^a(y|v) := \langle \widehat{\mu}_{Y^a|V}(v), \phi(y)\rangle_{\mathcal{H}_\mathcal{Y}}$ at 1000 equally spaced points $y \in [-20, 20]$. We set the negative values to zero and renormalize the result to integrate to one over the grid. Replacing $F_{Y^a|v}$ by the cumulative distribution function $\widehat{F}_{Y^a|v}$ of the renormalized density yields the plug-in estimators $\widehat{\mathrm{CVaR}}_\alpha(Y^a|v)$ and $\widehat{\mathrm{CVaR}}_\alpha(Y^1|v) - \widehat{\mathrm{CVaR}}_\alpha(Y^0|v)$.

We compare the estimates of the $\mathrm{CVaR}_{0.1}$ treatment effect with its true values in Figure 2(a). As we can see, the median CCME-DR estimates are close to the true values across all v. Because of the confounding by U, the median One-Step estimates are larger than the true values at every v. Since the CPME embedding does not depend on v, the CPME estimates are constant across all values of v. We also report the mean absolute errors (MAEs) in Figure 2(b). For the CVaR treatment effect, the MAE of CCME-DR is substantially smaller than those of One-Step and CPME. This is not surprising, as One-Step only adjusts for V and not U. For the CATE, the MAEs of CCME-DR, CPME, and CATE-DR are all significantly smaller than that of One-Step, which reflects how these three estimators handle the confounder U through the pseudo-outcomes.

Task 2: conditional two-sample test. We test the null hypothesis H_0 of Appendix E with the statistic $\widehat{T}$ as defined in (20). With $\widehat{\mathcal{V}}$ as the set of 100 equally spaced points in $[-2, 2]$, we denote the average of $\widehat{T}$ over $\widehat{\mathcal{V}}$ by $\overline{T} := \frac{1}{|\widehat{\mathcal{V}}|}\sum_{v\in\widehat{\mathcal{V}}}\widehat{T}(v)$.

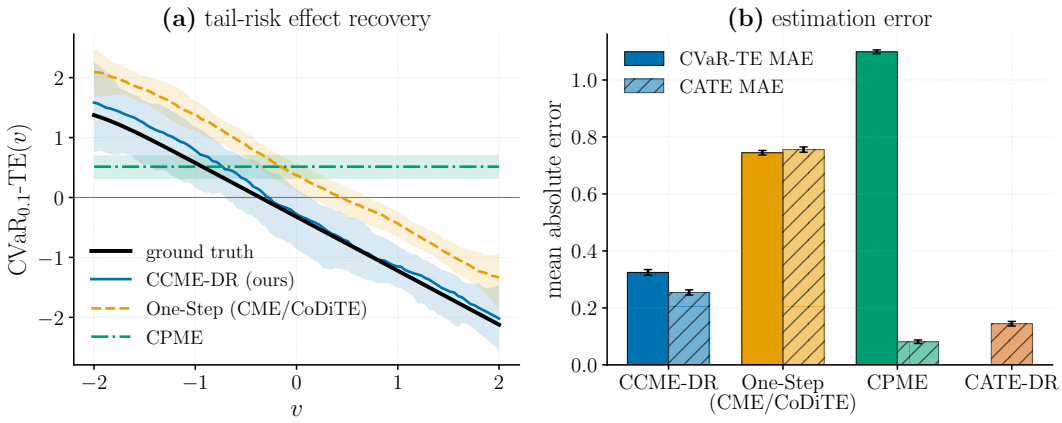


Figure 2. CVaR treatment effects (Task 1). (a) The estimated $\mathrm{CVaR}_{0.1}$ treatment effect against v: median over 100 replications. The shaded regions show pointwise intervals between the 10th and 90th percentiles. (b) MAEs of the $\mathrm{CVaR}_{0.1}$ treatment effect and the CATE, with standard errors over the replications. CVaR-TE denotes the CVaR treatment effect. Since CATE-DR estimates only the CATE, this estimator has no CVaR-TE bar.

For the Ridge Regression estimator, $\widehat{\Delta}^{(1)}(v)$ has the form

$$\widehat{\Delta}^{(1)}(v) \coloneqq \sum_{i \in \mathcal{I}_1} \beta_i^{(1)}(v)\, \widehat{\mathcal{C}}(Z_{1i}), \qquad \widehat{\mathcal{C}}(z) \coloneqq \widehat{\xi}^1(z) - \widehat{\xi}^0(z),$$
$$\left(\beta_i^{(1)}(v)\right)_{i \in \mathcal{I}_1} \coloneqq \left(\mathbf{K}_V^{(1)} + \tfrac{n}{2}\lambda_1 \mathbf{I}_{n/2}\right)^{-1} \mathbf{k}_V^{(1)}(v),$$

where $\mathbf{K}_V^{(1)}$ and $\mathbf{k}_V^{(1)}(v)$ are $\mathbf{K}_V$ and $\mathbf{k}_V(v)$ as in (5), computed on $\mathcal{I}_1$, and $\widehat{\xi}^1$ and $\widehat{\xi}^0$ are the pseudo-outcomes (3) for Y^1 and Y^0, with the nuisance functions estimated on $\mathcal{D}_0$. The difference $\widehat{\Delta}^{(2)}(v)$ has the same form with $\mathcal{I}_2$. Thus, we can write:

$$\overline{T} = \sum_{i \in \mathcal{I}_1} f_i, \qquad \text{where } f_i \coloneqq \frac{1}{|\widehat{\mathcal{V}}|} \sum_{v \in \widehat{\mathcal{V}}} \beta_i^{(1)}(v) \langle \widehat{\mathcal{C}}(Z_{1i}), \widehat{\Delta}^{(2)}(v) \rangle_{\mathcal{H}_\mathcal{Y}}.$$

The statistics for One-Step, CPME, and CATE-DR are calculated similarly with their own versions of $\beta_i^{(1)}(v)$ and $\widehat{\mathcal{C}}$. Let $\bar{f}$ and s_f be the mean and the standard deviation of $(f_i)_{i \in \mathcal{I}_1}$. Our test statistic is $T^\dagger \coloneqq \sqrt{n/2}\, \bar{f}/s_f$, and we reject the null hypothesis at level $\alpha = 0.05$ when $T^\dagger$ is at least the 0.95 quantile of the standard normal distribution.

We compute the CCME-DR, One-Step, and CPME statistics with the Ridge Regression estimator, $\frac{n}{2}\lambda_1 = m\lambda_0 = 1$ with $n = 1000$. We consider four scenarios, which have the same distributions of V, U, X, $\pi(X)$, A, and S as Tasks 1 and 3 and differ only in the potential outcomes:

$$\begin{aligned}
&\text{Scenario 1:} && Y^0 = V + 0.8U + \epsilon_0, && Y^1 = V + 0.8U + \epsilon_1,\\
&\text{Scenario 2:} && Y^0 = V - 0.8U + \epsilon_0, && Y^1 = V + 0.8U + \epsilon_1,\\
&\text{Scenario 3:} && Y^0 = 0.8U + 3S\, w(V) + \epsilon_0, && Y^1 = 0.8U + 3S\,(1 - w(V)) + \epsilon_1,\\
&\text{Scenario 4:} && Y^0 = V + 0.8U + \epsilon_0, && Y^1 = V + S\,\delta(V) + 0.8U + \epsilon_1,
\end{aligned}$$

where $w(v) \coloneqq \mathsf{sigmoid}(2v)$. The noise terms ϵ_0, ϵ_1 follow $\mathcal{N}(0, 0.7^2)$ in Scenarios 1, 2, and 4 and $\mathcal{N}(0, 0.6^2)$ in Scenario 3. In all four scenarios, $\mathbb{E}[Y^1 - Y^0 | V] = 0$. The null hypothesis H_0 holds in Scenarios 1 and 2. In Scenario 2, $P(Y^1|V) = P(Y^0|V)$, while $P(Y^1|X) \neq P(Y^0|X)$. In Scenario 3, $P(Y^1|V) \neq P(Y^0|V)$, while $P(Y^1) = P(Y^0)$. In Scenario 4, $P(Y^1) \neq P(Y^0)$ and $P(Y^1|V) \neq P(Y^0|V)$.

We report the rejection rates in Table 3. In Scenarios 1 and 2, the type I error rates of the CCME-DR, CPME, and CATE-DR tests are close to the nominal level. In Scenarios 3 and 4, the CCME-DR and One-Step tests reject H_0 in every replication. However, because of the confounding by U, the type I error rate of the One-Step test in Scenarios 1 and 2 is significantly larger than the nominal level. Since the CPME statistic compares the marginal distributions of Y^1 and Y^0, which coincide in Scenario 3,

the CPME test has trivial power in this scenario. Similarly, since $\mathbb{E}[Y^1 - Y^0|V] = 0$ in Scenarios 3 and 4, the CATE-DR test has almost no power in these two scenarios. Thus, the CCME-DR test is the only one of the four tests whose type I error rate is close to the nominal level in Scenarios 1 and 2 and that rejects H_0 in every replication of Scenarios 3 and 4.

Table 3. Rejection rates of the conditional two-sample tests at level $\alpha = 0.05$ (Task 2), with standard errors, over 2000 replications in each scenario, with sample size 2000. The target level is α in Scenarios 1 and 2, where H_0 holds, and 1 in Scenarios 3 and 4, where H_0 fails. In each scenario, the bold entries are those whose 95% confidence intervals cover the target level.

Test	Scenario 1	Scenario 2	Scenario 3	Scenario 4
CCME-DR	$\mathbf{0.049 \pm 0.005}$	$\mathbf{0.052 \pm 0.005}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
One-Step [PSSM21]	0.999 ± 0.001	0.140 ± 0.008	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
CPME [ZBG25]	$\mathbf{0.048 \pm 0.005}$	$\mathbf{0.056 \pm 0.005}$	0.040 ± 0.004	$\mathbf{1.000 \pm 0.000}$
CATE-DR	$\mathbf{0.053 \pm 0.005}$	$\mathbf{0.048 \pm 0.005}$	0.051 ± 0.005	0.041 ± 0.004

Task 3: risk-sensitive policy learning. A policy is a map $d : \mathcal{V} \to \{0, 1\}$. The tail-sensitive value of a policy d is

$$\mathrm{Val}_\alpha(d) := \mathbb{E}\Big[\mathrm{CVaR}_\alpha\Big(Y^{d(V)}|V\Big)\Big].$$

The policy $d^\star_\alpha(v) := \mathbb{I}(\mathrm{CVaR}_\alpha(Y^1|v) > \mathrm{CVaR}_\alpha(Y^0|v))$ maximizes Val_α. This optimal policy depends on the lower tails of the conditional counterfactual distributions, which the CATE does not determine. For each of CCME-DR, One-Step, and CPME, we use the plug-in policy $\widehat{d}_\alpha(v) := \mathbb{I}(\widehat{\mathrm{CVaR}}_\alpha(Y^1|v) > \widehat{\mathrm{CVaR}}_\alpha(Y^0|v))$, with the two CVaRs computed from the estimated densities as in Task 1. The CATE-DR policy is $\widehat{d}(v) := \mathbb{I}(\widehat{\mathbb{E}}[Y^1 - Y^0|V = v] > 0)$ with the CATE estimated by CATE-DR. This policy does not depend on α. We report the regret $\mathrm{Val}_\alpha(d^\star_\alpha) - \mathrm{Val}_\alpha(\widehat{d}_\alpha)$ (and $\mathrm{Val}_\alpha(d^\star_\alpha) - \mathrm{Val}_\alpha(\widehat{d})$ for the CATE-DR policy), evaluated under the true distributions, for $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}$ in Figure 3. The figure shows that, across the values of α, the median regret of CCME-DR is close to zero and an order of magnitude smaller than that of One-Step. In contrast, the CPME and CATE-DR policies treat every unit. As a result, they have the largest regret, which increases as α decreases.

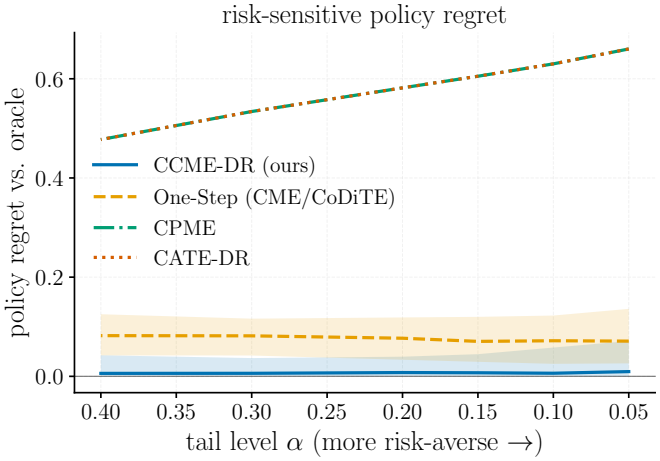


Figure 3. Regret of the estimated risk-sensitive policies relative to the optimal policy $d^\star_\alpha$ (Task 3), as a function of the tail level α, which decreases to the right. The lines are the medians over 100 replications. The shaded regions are pointwise intervals between the 10th and 90th percentiles.

H Background on Function Spaces and RKHS

We provide a brief review of the Bochner integral, (fractional) Sobolev spaces, and reproducing kernel Hilbert spaces (RKHS). A comprehensive treatment of these topics can be found in [HvNVW16, Chapters 1 and 2].

H.1 Function Spaces and Fourier Analysis

In this subsection, let $\Omega \subseteq \mathbb{R}^d$ denote a generic domain (which will correspond to either $\mathcal{Y}$ or $\mathcal{V}$ in the main paper).

Definition 1 (Lebesgue Space $L^p(\Omega)$)**.** *A function $f : \Omega \to \mathbb{R}$ is said to be in the* Lebesgue space *$L^p(\Omega)$ for $1 \leq p \leq \infty$ if it is measurable and its L^p norm is finite. That is,*

$$\|f\|_{L^p(\Omega)} \coloneqq \left(\int_\Omega |f(x)|^p \, \mathrm{d}x \right)^{1/p} < \infty,$$

and for $p = \infty$,

$$\|f\|_{L^\infty(\Omega)} \coloneqq \inf\{r \geq 0 : \mathfrak{m}(\{|f(x)| > r\}) = 0\} < \infty,$$

where $\mathfrak{m}$ denotes the Lebesgue measure on $\mathbb{R}^d$.

Definition 2 (Weak Derivatives)**.** *Let f be a function in $L^1(\Omega)$. A function $g \in L^1(\Omega)$ is the* weak derivative *of f with respect to x_i if for all smooth test functions with compact support $\zeta \in C_c^\infty(\Omega)$, the following identity holds:*

$$\int_\Omega f(x) \frac{\partial \zeta(x)}{\partial x_i} \, \mathrm{d}x = -\int_\Omega g(x)\zeta(x) \, \mathrm{d}x.$$

*For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$, the αth-*weak derivative*, denoted by $D^\alpha f$, must satisfy the following identity for all $\zeta \in C_c^\infty(\Omega)$:*

$$\int_\Omega f(x)(D^\alpha \zeta(x)) \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega (D^\alpha f(x))\zeta(x) \, \mathrm{d}x.$$

Definition 3 (Sobolev Space $W^{k,p}(\Omega)$)**.** *For $k \in \mathbb{N}_0$ and $p \in [1, \infty]$, the* Sobolev space *$W^{k,p}(\Omega)$ is the set of all functions $f \in L^p(\Omega)$ such that for every multi-index $\alpha \in \mathbb{N}_0^d$ with total order $|\alpha| \leq k$, the weak derivative $D^\alpha f$ exists and is also in $L^p(\Omega)$.*

The norm for this space is defined as:

$$\|f\|_{W^{k,p}(\Omega)} \coloneqq \sum_{|\alpha| \leq k} \|D^\alpha f\|_{L^p(\Omega)}.$$

Definition 4 (Fourier Transform)**.** *Let $f \in L^1(\mathbb{R}^d)$ be an integrable function. The Fourier transform of f, denoted $\widehat{f}$ or $\mathcal{F}[f]$, is defined by the integral:*

$$\widehat{f}(\omega) \coloneqq \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\omega^\top x} f(x) \, \mathrm{d}x,$$

where $\omega \in \mathbb{R}^d$ is the frequency variable.

One important property of the Fourier transform is that it is a linear isometry on $L^2(\mathbb{R}^d)$ as the following theorem shows:

Theorem 5 (Plancherel's Theorem)**.** *The Fourier transform, initially defined on $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, extends uniquely to a bounded linear operator on $L^2(\mathbb{R}^d)$. Moreover, the operator is unitary, meaning that for any $f \in L^2(\mathbb{R}^d)$:*

$$\|f\|_{L^2(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} |f(x)|^2 \, \mathrm{d}x = \int_{\mathbb{R}^d} |\widehat{f}(\omega)|^2 \, \mathrm{d}\omega = \|\widehat{f}\|_{L^2(\mathbb{R}^d)}^2.$$

While $W^{k,p}$ spaces are defined for integer k, it is often necessary to consider spaces of functions with fractional smoothness $s \in \mathbb{R}$. The Fourier transform provides a natural characterization of these spaces when $p = 2$.

Definition 5 (Fractional Sobolev Space $H^s(\mathbb{R}^d)$)**.** *For any $s \in \mathbb{R}$, the fractional Sobolev space $H^s(\mathbb{R}^d)$ is defined as the space of tempered distributions f such that:*

$$\|f\|_{H^s(\mathbb{R}^d)}^2 \coloneqq \int_{\mathbb{R}^d} (1 + \|\omega\|^2)^s |\widehat{f}(\omega)|^2 \, \mathrm{d}\omega < \infty.$$

When $s = k$ is a non-negative integer, this space is isomorphic to the Sobolev space $W^{k,2}(\mathbb{R}^d)$ equipped with the standard norm.

For a domain $\Omega \subset \mathbb{R}^d$ and a non-integer $s > 0$, we define $W^{s,2}(\Omega) \coloneqq \{F|_\Omega : F \in H^s(\mathbb{R}^d)\}$ with the norm $\|f\|_{W^{s,2}(\Omega)} \coloneqq \inf\{\|F\|_{H^s(\mathbb{R}^d)} : F \in H^s(\mathbb{R}^d),\ F|_\Omega = f\}$.

A crucial property of Sobolev spaces is the embedding theorem, which relates regularity to integrability; it allows us to embed a space with higher derivatives into a space with higher integrability. To state the theorem, we introduce two notions of smooth domains.

Definition 6 (Lipschitz Boundary)**.** *The domain $\Omega \subset \mathbb{R}^d$ is said to have a* Lipschitz boundary *if it is a non-empty, bounded set and its boundary $\partial\Omega$ satisfies the following property: For every point $x \in \partial\Omega$, there exists a radius $r > 0$ and an orthogonal coordinate system $(y_1, \ldots, y_d)$ (which may depend on x) such that if we define the cylinder $C \coloneqq \{y : \|(y_1, \ldots, y_{d-1})\| < r, |y_d| < r\}$, then:*

$$\Omega \cap C = \{y \in C : y_d < \gamma(y_1, \ldots, y_{d-1})\},$$

where $\gamma : \mathbb{R}^{d-1} \to \mathbb{R}$ is a Lipschitz continuous function.

For an integer $s > d/2$ and a domain Ω with a Lipschitz boundary, the space $\{F|_\Omega : F \in H^s(\mathbb{R}^d)\}$ with the infimum norm coincides with the space $W^{s,2}(\Omega)$ of Definition 3, with an equivalent norm [Wen04, Section 10.7].

Definition 7 (Uniform Interior Cone Condition)**.** *The domain Ω is said to satisfy a* uniform interior cone condition *if there exist an angle $\theta \in (0, \pi/2)$ and a radius $\delta > 0$ such that for every $x \in \bar{\Omega}$, there exists a unit vector $\xi \in \mathbb{R}^d$ such that the cone defined by:*

$$\mathfrak{C}(x, \xi, \theta, \delta) \coloneqq \left\{x + z : z \in \mathbb{R}^d, \|z\| < \delta, z^\top \xi \geq \|z\| \cos\theta\right\}$$

is contained in Ω.

Theorem 6 (Sobolev Embedding)**.** *Suppose that either $\Omega = \mathbb{R}^d$ or Ω is a bounded set with a Lipschitz boundary in $\mathbb{R}^d$. Let $(s, p), (t, q) \in \mathbb{N}_0 \times [1, \infty) \cup [0, \infty) \times \{2\}$ with $p \leq q$. If $s - \frac{d}{p} = t - \frac{d}{q}$, then we have the continuous embedding:*

$$W^{s,p}(\Omega) \hookrightarrow W^{t,q}(\Omega).$$

Remark 3 (Boundary Regularity)**.** *The embeddings stated above do not require any boundary condition when $\Omega = \mathbb{R}^d$. However, if the domain is restricted to a proper open subset $\Omega \subset \mathbb{R}^d$, these embeddings require Ω to satisfy specific regularity conditions. Typically, Ω must have a* Lipschitz boundary *(or satisfy the strong cone property) to ensure the existence of a bounded extension operator from $W^{s,p}(\Omega)$ to $W^{s,p}(\mathbb{R}^d)$.*

H.2 RKHS and Bochner's Theorem

Definition 8 (Reproducing Kernel Hilbert Space)**.** *A* Reproducing Kernel Hilbert Space *(RKHS) $\mathcal{H}_\mathcal{Y} \subseteq (\mathcal{Y} \to \mathbb{R})$ is a Hilbert space in which the point evaluation functional is a bounded linear operator. That is, for every $y \in \mathcal{Y}$, there exists a constant $C_y > 0$ such that $|f(y)| \leq C_y \|f\|_{\mathcal{H}_\mathcal{Y}}$ for all $f \in \mathcal{H}_\mathcal{Y}$.*

By the Riesz representation theorem, the boundedness of the evaluation functional implies the existence of a unique function $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, called the *reproducing kernel*. For every $y \in \mathcal{Y}$, the function $k(\cdot, y)$ belongs to $\mathcal{H}_\mathcal{Y}$. We define the *canonical feature map* $\phi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ as $\phi(y) \coloneqq k(\cdot, y)$.

For all $f \in \mathcal{H}_\mathcal{Y}$ and $y \in \mathcal{Y}$, the function value is given by the inner product:

$$f(y) = \langle f, \phi(y)\rangle_{\mathcal{H}_\mathcal{Y}} = \langle f, k(\cdot, y)\rangle_{\mathcal{H}_\mathcal{Y}}.$$

Definition 9 (Positive Definite Kernel)**.** *A function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ is called a* positive definite kernel *if for any integer $N \geq 1$, any set of points $\{y_1, \ldots, y_N\} \subset \mathbb{R}^d$, and any coefficients $\{c_1, \ldots, c_N\} \subset \mathbb{C}$, the following inequality holds:*

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i \overline{c_j} k(y_i, y_j) \geq 0.$$

We call k *strictly positive definite* if, in addition, the double sum is positive whenever the points $y_1, \ldots, y_N$ are distinct and $(c_1, \ldots, c_N) \neq 0$.

For any finite set of distinct points $\{y_1, \ldots, y_M\} \subset \mathcal{Y}$, the *Gram matrix* (or kernel matrix) $\mathbf{K}_M \in \mathbb{R}^{M \times M}$ with entries $(\mathbf{K}_M)_{ij} = k(y_i, y_j) = \langle \phi(y_i), \phi(y_j)\rangle_{\mathcal{H}_\mathcal{Y}}$ is invertible when k is strictly positive definite, as stated in the following proposition:

Proposition 2 (Independence of Feature Maps)**.** *If the kernel k is strictly positive definite, then for any set of distinct points $\{y_1, \dots, y_M\} \subset \mathcal{Y}$, the Gram matrix $\mathbf{K}_M$ is invertible and the set of feature vectors $\{\phi(y_1), \dots, \phi(y_M)\}$ is linearly independent in $\mathcal{H}_\mathcal{Y}$.*

A function $f : \mathcal{V} \to \mathcal{H}_\mathcal{Y}$ is called a *simple function* if it can be written as $f(v) = \sum_{i=1}^N x_i \mathbb{I}_{E_i}(v)$, where $\mathbb{I}_{E_i}(\cdot)$ is the indicator function and $\{E_i\}_{i=1}^N$ are disjoint, Lebesgue measurable subsets of $\mathcal{V}$, and $\{x_i\}_{i=1}^N$ are vectors in $\mathcal{H}_\mathcal{Y}$. The integral of $f(v)$ is given by $\int_\mathcal{V} f(v)\,\mathrm{d}v \coloneqq \sum_{i=1}^N x_i \mathfrak{m}(E_i)$. Note that the result of this integral is a single vector in the Hilbert space $\mathcal{H}_\mathcal{Y}$.

Definition 10 (Bochner Integral)**.** *A function $h : \mathcal{V} \to \mathcal{H}_\mathcal{Y}$ is said to be* Bochner integrable *if it satisfies two conditions:*

1. *There exists a sequence of simple functions $\{f_n\}_{n=1}^\infty$ such that $\lim_{n\to\infty} f_n(v) = h(v)$ for almost every $v \in \mathcal{V}$. In this case, we say that h is* Bochner measurable*.*
2. *The real-valued function $v \mapsto \|h(v)\|_{\mathcal{H}_\mathcal{Y}}$ is Lebesgue integrable, i.e.,*

$$\int_\mathcal{V} \|h(v)\|_{\mathcal{H}_\mathcal{Y}}\,\mathrm{d}v < \infty.$$

If a function h is Bochner integrable, its Bochner integral *is defined as the limit of the integrals of its approximating sequence of simple functions:*

$$\int_\mathcal{V} h(v)\,\mathrm{d}v \coloneqq \lim_{n\to\infty} \int_\mathcal{V} f_n(v)\,\mathrm{d}v.$$

The completeness of the Hilbert space $\mathcal{H}_\mathcal{Y}$ guarantees that this limit exists and is unique (up to sets of Lebesgue measure zero), regardless of the choice of the approximating sequence $\{f_n\}$.

Of particular interest in this work are kernels that are translation-invariant. Such kernels can be written as $k(y, y') = \varphi(y - y')$ for some continuous function $\varphi : \mathbb{R}^d \to \mathbb{C}$. In this setting, Bochner's theorem provides a connection between positive definite functions and Fourier analysis.

Theorem 7 (Bochner's Theorem)**.** *A continuous function $\varphi : \mathbb{R}^d \to \mathbb{C}$ is positive definite (i.e., it defines a positive definite kernel $k(y, y') = \varphi(y - y')$) if and only if it is the Fourier transform of a non-negative, finite Borel measure on $\mathbb{R}^d$.*

In our case, we assume this measure has a density, which we call $\widehat{\varphi}(\omega)$. This means $\varphi(z) = \mathcal{F}^{-1}[\widehat{\varphi}](z)$, and we require $\widehat{\varphi}(\omega) \geq 0$ for all ω. This spectral perspective allows us to express the RKHS norm via an integral in the frequency domain.

Proposition 3 (RKHS Norm in the Fourier Domain)**.** *For a translation-invariant kernel $k(y, y') = \varphi(y - y')$ on $\mathbb{R}^{d_y}$ with the RKHS $\mathcal{H}_{\mathbb{R}^{d_y}}$, the RKHS norm of a function $f \in \mathcal{H}_{\mathbb{R}^{d_y}}$ is given by:*

$$\|f\|^2_{\mathcal{H}_{\mathbb{R}^{d_y}}} = \int_{\mathbb{R}^{d_y}} \frac{|\widehat{f}(\omega)|^2}{\widehat{\varphi}(\omega)}\,\mathrm{d}\omega.$$

For a set $\mathcal{Y} \subset \mathbb{R}^{d_y}$, the RKHS $\mathcal{H}_\mathcal{Y}$ of the restricted kernel satisfies $\mathcal{H}_\mathcal{Y} = \{F|_\mathcal{Y} : F \in \mathcal{H}_{\mathbb{R}^{d_y}}\}$ and $\|f\|_{\mathcal{H}_\mathcal{Y}} = \inf\{\|F\|_{\mathcal{H}_{\mathbb{R}^{d_y}}} : F \in \mathcal{H}_{\mathbb{R}^{d_y}},\ F|_\mathcal{Y} = f\}$ for every $f \in \mathcal{H}_\mathcal{Y}$.

H.3 Operators on Hilbert Spaces

Let $\mathcal{U}$ and $\mathcal{W}$ be separable Hilbert spaces, and let $(e_k)_{k\geq 1}$ be an orthonormal basis of $\mathcal{U}$. For a bounded linear operator $T : \mathcal{U} \to \mathcal{W}$, we denote by $\|T\|_{\text{op}}$ and $\|T\|_{\text{HS}}$ the operator norm and the Hilbert-Schmidt norm of T:

$$\|T\|_{\text{op}} \coloneqq \sup_{\|x\|_\mathcal{U} \leq 1} \|Tx\|_\mathcal{W}, \qquad \|T\|^2_{\text{HS}} \coloneqq \sum_{k\geq 1} \|Te_k\|^2_\mathcal{W}.$$

The Hilbert-Schmidt inner product of two such operators S and T is $\langle S, T\rangle_{\text{HS}} \coloneqq \sum_{k\geq 1} \langle Se_k, Te_k\rangle_\mathcal{W}$. For vectors in $\mathbb{R}^M$, $\|\cdot\|$ is the Euclidean norm. For matrices, $\|\cdot\|_{\text{op}}$ is the spectral norm. Moreover, the Hilbert-Schmidt norm and inner product of matrices are the Frobenius norm $\|\cdot\|_F$ and the Frobenius inner product. We denote by T^* the adjoint of T: the operator from $\mathcal{W}$ to $\mathcal{U}$ with

$\langle Tx, y\rangle_{\mathcal{W}} = \langle x, T^*y\rangle_{\mathcal{U}}$ for all $x \in \mathcal{U}$ and $y \in \mathcal{W}$. An operator T on $\mathcal{U}$ is self-adjoint if $T^* = T$ and positive if $\langle Tx, x\rangle_{\mathcal{U}} \geq 0$ for all $x \in \mathcal{U}$. For a positive self-adjoint operator T on $\mathcal{U}$, we denote by $\mathrm{Tr}(T)$ the trace of T: $\mathrm{Tr}(T) := \sum_{k\geq 1}\langle Te_k, e_k\rangle_{\mathcal{U}}$. These sums do not depend on the basis. For $a \in \mathcal{W}$ and $b \in \mathcal{U}$, we denote by $a \otimes b$ the operator $x \mapsto \langle b, x\rangle_{\mathcal{U}} a$ from $\mathcal{U}$ to $\mathcal{W}$, which satisfies $\|a \otimes b\|_{\mathrm{HS}} = \|a\|_{\mathcal{W}}\|b\|_{\mathcal{U}}$ and $\langle T, a \otimes b\rangle_{\mathrm{HS}} = \langle Tb, a\rangle_{\mathcal{W}}$.

I Proof of the Identification of CCME

In this section, we provide a proof of the identification of CCME presented in Proposition 1. We begin by formalizing the definition of V.

Let $(\Omega, \mathcal{F}_Z, \mathbb{P})$ be the probability space on which the random variables $Z = (X, A, Y)$ are defined. Let $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{V}, \Sigma_{\mathcal{V}})$ denote the measurable spaces associated with X and V, respectively. We assume that the mapping $\eta : \mathcal{X} \to \mathcal{V}$ defining $V = \eta(X)$ is $(\Sigma_{\mathcal{X}}, \Sigma_{\mathcal{V}})$-measurable. Consequently, the σ-algebra generated by V, denoted $\sigma(V) := \{X^{-1}(\eta^{-1}(E)) : E \in \Sigma_{\mathcal{V}}\}$, satisfies $\sigma(V) \subseteq \sigma(X) \subseteq \mathcal{F}_Z$.

Proof of Proposition 1. We first verify that the random variable $\xi(Z)$ is Bochner integrable by showing it satisfies the two conditions of Definition 10.

We begin with the feature map $\phi(Y) = k_{\mathcal{Y}}(\cdot, Y)$. Since $\mathcal{Y}$ is a subset of a Euclidean space and the kernel $k_{\mathcal{Y}}$ is continuous, $\phi(Y)$ is continuous as a mapping from $\mathcal{Y}$ to $\mathcal{H}_{\mathcal{Y}}$. Continuous functions on separable domains can be approximated by simple functions (e.g., piecewise constant functions on a grid); thus, $\phi(Y)$ is Bochner measurable.

Next, consider the term $\mu_0(X)$. By Assumption 2, the kernel is bounded: $\sup_{y\in\mathcal{Y}} k_{\mathcal{Y}}(y, y) \leq B_k$ for some constant $B_k > 0$. Since $\phi(Y)$ is Bochner measurable and bounded, it is Bochner integrable. Since the conditional expectation of a Bochner integrable random variable is defined to be a Bochner measurable function [PM20, Definition 2.5], $\mu_0(X)$ is Bochner measurable. Since $\xi(Z)$ is a linear combination of $\phi(Y)$ and $\mu_0(X)$, it is also Bochner measurable.

We verify that $\mathbb{E}[\|\xi(Z)\|_{\mathcal{H}_{\mathcal{Y}}}] < \infty$. By the triangle inequality:

$$\|\xi(Z)\|_{\mathcal{H}_{\mathcal{Y}}} \leq \frac{|A|}{\pi(X)}\|\phi(Y)\|_{\mathcal{H}_{\mathcal{Y}}} + \left|1 - \frac{A}{\pi(X)}\right|\|\mu_0(X)\|_{\mathcal{H}_{\mathcal{Y}}}.$$

By Jensen's inequality for Bochner integrals, the norm of the conditional mean is bounded by the conditional expectation of the norm:

$$\|\mu_0(X)\|_{\mathcal{H}_{\mathcal{Y}}} = \|\mathbb{E}[\phi(Y)|X, A = 1]\|_{\mathcal{H}_{\mathcal{Y}}} \leq \mathbb{E}[\|\phi(Y)\|_{\mathcal{H}_{\mathcal{Y}}}|X, A = 1] \leq B_k^{1/2}.$$

Invoking the strong positivity assumption ($\pi(X) \geq \varepsilon > 0$), we verify the bound almost surely:

$$\|\xi(Z)\|_{\mathcal{H}_{\mathcal{Y}}} \leq \frac{1}{\varepsilon}B_k^{1/2} + \left(1 + \frac{1}{\varepsilon}\right)B_k^{1/2} = B_k^{1/2}(2\varepsilon^{-1} + 1) < \infty.$$

Since the norm is uniformly bounded, its expectation is finite. Thus, $\xi(Z)$ is Bochner integrable.

We now prove (2). By the law of iterated expectations, since $\sigma(V) \subseteq \sigma(X)$, we have:

$$\mathbb{E}[\xi(Z)|V] = \mathbb{E}\big[\mathbb{E}[\xi(Z)|X] \mid V\big]. \tag{23}$$

We first compute the inner conditional expectation $\mathbb{E}[\xi(Z)|X]$. Since $\pi(X)$ and $\mu_0(X)$ are $\sigma(X)$-measurable, they can be factored out:

$$\begin{aligned}\mathbb{E}[\xi(Z)|X] &= \mathbb{E}\left[\frac{A}{\pi(X)}\phi(Y) - \frac{A}{\pi(X)}\mu_0(X) + \mu_0(X) \;\middle|\; X\right] \\ &= \frac{1}{\pi(X)}\mathbb{E}[A\phi(Y)|X] - \frac{\mu_0(X)}{\pi(X)}\mathbb{E}[A|X] + \mu_0(X). \end{aligned} \tag{24}$$

Recall that $\mathbb{E}[A|X] = P(A = 1|X) = \pi(X)$. For the term $\mathbb{E}[A\phi(Y)|X]$, we apply the conditional ignorability assumption ($Y^1 \perp A|X$):

$$\begin{aligned}\mathbb{E}[A\phi(Y)|X] &= \mathbb{E}[A\phi(Y^1)|X] \\ &= \mathbb{E}[\phi(Y^1)|X, A = 1]P(A = 1|X) \\ &= \mathbb{E}[\phi(Y^1)|X]\pi(X).\end{aligned}$$

We also note that $\mu_0(X) = \mathbb{E}[\phi(Y)|X, A=1] = \mathbb{E}[\phi(Y^1)|X, A=1] = \mathbb{E}[\phi(Y^1)|X]$. Substituting these results back into (24) yields:

$$\begin{aligned}\mathbb{E}[\xi(Z)|X] &= \frac{1}{\pi(X)}\big(\mathbb{E}[\phi(Y^1)|X]\pi(X)\big) - \frac{\mathbb{E}[\phi(Y^1)|X]}{\pi(X)}\pi(X) + \mathbb{E}[\phi(Y^1)|X] \\ &= \mathbb{E}[\phi(Y^1)|X] - \mathbb{E}[\phi(Y^1)|X] + \mathbb{E}[\phi(Y^1)|X] \\ &= \mathbb{E}[\phi(Y^1)|X].\end{aligned}$$

Finally, substituting this result into (23) and using the law of iterated expectations again yields:

$$\begin{aligned}\mathbb{E}[\xi(Z)|V] &= \mathbb{E}\big[\mathbb{E}[\phi(Y^1)|X] \mid V\big] \\ &= \mathbb{E}[\phi(Y^1)|V] \\ &= \mu_{Y^1|V}(V).\end{aligned}$$

This concludes the proof. □

J Proof of the Upper Bound for the Meta-Estimator (Theorem 1)

For brevity, we denote by $\widetilde{\mu}_{Y^1|V}(v)$ the conditional mean embedding of the pseudo-outcome, i.e., $\widetilde{\mu}_{Y^1|V}(v) \coloneqq \mathbb{E}[\widehat{\xi}(Z)|V=v, \mathcal{D}_0]$. We split the error into two terms:

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\big] &\leq 2\mathbb{E}\big[\|\widehat{\mu}_{Y^1|V}(V) - \widetilde{\mu}_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\big] \\ &\quad + 2\mathbb{E}\big[\|\widetilde{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\big] \\ &= 2\mathcal{R}^2_{\widehat{\xi}}(\widehat{\mu}_{Y^1|V}) + 2\mathbb{E}\big[\|\widetilde{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\big].\end{aligned} \tag{25}$$

To bound the second term on the right-hand side, we use the following lemma, the proof of which is provided in Appendix O:

Lemma 1 (Nuisance Error Bound)**.** *Let Assumptions 1 and 2 hold. In addition, assume that $\widehat{\pi}(x), \pi(x) \in [\varepsilon, 1)$ for all $x \in \mathcal{X}$ and $\sup_{x\in\mathcal{X}}\|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$ for some constant $B_{\mu_0} > 0$. Then we have the following bound:*

$$\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \lesssim \min\big\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\big\}. \tag{26}$$

With this lemma, we insert (26) back in the right-hand side of (25) to finish the proof. ■

K Proof of the Upper Bound for the Ridge Regression Estimator (Theorem 2)

K.1 Notations and Setup

We denote $\lambda \coloneqq \lambda_1$, $\mu \coloneqq \mu_{Y^1|V}$, $\widetilde{\mu}_{Y^1|V}(v) \coloneqq \mathbb{E}[\widehat{\xi}(Z)|V=v, \mathcal{D}_0]$, and $\|h\|^2_{L_2(P_V)} \coloneqq \int_\mathcal{V}\|h(v)\|^2_{\mathcal{H}_\mathcal{Y}}\, p_V(v)\,\mathrm{d}v$ for $h \in (\mathcal{V} \to \mathcal{H}_\mathcal{Y})$, where P_V is the distribution of V and p_V is its density in Assumption 4.

For $u = (u_1, \ldots, u_n) \in \mathcal{H}^n_\mathcal{Y}$, we denote:

$$\widehat{F}[u](v) \coloneqq \sum_{i=1}^n \beta_i(v)u_i, \qquad \boldsymbol{\beta}(v) \coloneqq (\mathbf{K}_V + n\lambda\mathbf{I}_n)^{-1}\mathbf{k}_V(v), \qquad v \in \mathcal{V}. \tag{27}$$

By (5), we have $\widehat{\mu}_{\mathrm{RR}} = \widehat{F}[u]$ with $u_i = \widehat{\xi}(Z_{1i})$.

K.2 Preliminary Lemma

The following lemma provides a bound on the norm of the pseudo-outcome:

Lemma 2. *Suppose that $\sup_{y\in\mathcal{Y}} k_\mathcal{Y}(y,y) \leq B_k$, that $\widehat{\pi}(x) \geq \varepsilon > 0$ for all $x \in \mathcal{X}$, and that $\sup_{x\in\mathcal{X}}\|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$. Then the pseudo-outcome satisfies $\|\widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \leq \sqrt{B_{\widehat{\xi}}}$ for all z, where $B_{\widehat{\xi}} \coloneqq (\sqrt{B_k} + B_{\mu_0})^2/\varepsilon^2$.*

Proof. We only write the computation for $z = Z_{1i}$, since the computation is the same for every z. Recall that

$$\widehat{\xi}(Z_{1i}) = \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\phi(Y_{1i}) + \left(1 - \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\right)\widehat{\mu}_0(X_{1i}).$$

We have:

$$\begin{aligned}\langle\widehat{\xi}(Z_{1i}), \widehat{\xi}(Z_{1i})\rangle_{\mathcal{H}_\mathcal{Y}} = \frac{A_{1i}^2}{\widehat{\pi}(X_{1i})^2}k_\mathcal{Y}(Y_{1i}, Y_{1i}) \\ + 2\frac{A_{1i}}{\widehat{\pi}(X_{1i})}\left(1 - \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\right)\langle\phi(Y_{1i}), \widehat{\mu}_0(X_{1i})\rangle_{\mathcal{H}_\mathcal{Y}} \\ + \left(1 - \frac{A_{1i}}{\widehat{\pi}(X_{1i})}\right)^2 \|\widehat{\mu}_0(X_{1i})\|^2_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

By assumption, $\|\widehat{\mu}_0(X_{1i})\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$. Since $A_{1i} \in \{0, 1\}$ and $\widehat{\pi}(X_{1i}) \geq \varepsilon$, the coefficients $A_{1i}/\widehat{\pi}(X_{1i})$ and $1 - A_{1i}/\widehat{\pi}(X_{1i})$ are at most $1/\varepsilon$ in absolute value. Using $k_\mathcal{Y}(Y_{1i}, Y_{1i}) \leq B_k$ and the Cauchy-Schwarz inequality for the cross term, we obtain:

$$\langle\widehat{\xi}(Z_{1i}), \widehat{\xi}(Z_{1i})\rangle_{\mathcal{H}_\mathcal{Y}} \leq \frac{B_k}{\varepsilon^2} + \frac{2\sqrt{B_k}B_{\mu_0}}{\varepsilon^2} + \frac{B_{\mu_0}^2}{\varepsilon^2} = \frac{(\sqrt{B_k} + B_{\mu_0})^2}{\varepsilon^2} =: B_{\widehat{\xi}}.$$

□

K.3 Proof of the Bound

We prove the bound in several steps. In Step 1, we bound $\|\widehat{F}[u]\|^2_{L_2(P_V)}$ in (27) for all $u \in \mathcal{H}_\mathcal{Y}^n$. In Step 2, we split the error of $\widehat{\mu}_{\mathrm{RR}}$ into four terms, which we bound in Steps 3–6.

Before Step 1, we introduce several notations and derive a couple of bounds that we will use throughout the proof. By Lemma 2, Jensen's inequality and $k_\mathcal{Y}(y, y) \leq B_k$, we have, for every z and $v \in \mathcal{V}$:

$$\|\widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \leq \sqrt{B_{\widehat{\xi}}}, \qquad \|\widetilde{\mu}_{Y^1|V}(v)\|_{\mathcal{H}_\mathcal{Y}} \leq \sqrt{B_{\widehat{\xi}}}, \qquad \|\mu(v)\|_{\mathcal{H}_\mathcal{Y}} \leq \sqrt{B_k}. \tag{28}$$

Denote $b := \widetilde{\mu}_{Y^1|V} - \mu$. Let (V, Z) be a new copy of (V_{11}, Z_{11}), independent of $\mathcal{D}_0$ and $\mathcal{D}_1$, and denote $U := \widehat{\xi}(Z) - b(V)$. On the sample, we denote $U_i := \widehat{\xi}(Z_{1i}) - b(V_{1i})$ for $i = 1, \ldots, n$. Since $\widehat{\xi}$ and b depend only on $\mathcal{D}_0$, the pairs (V_{1i}, U_i) are, conditionally on $\mathcal{D}_0$, independent copies of (V, U). Moreover, $\mathbb{E}[U|V, \mathcal{D}_0] = \widetilde{\mu}_{Y^1|V}(V) - b(V) = \mu(V)$. By (28), we have:

$$\begin{gathered}\sup_{v\in\mathcal{V}}\|b(v)\|_{\mathcal{H}_\mathcal{Y}} \leq \sqrt{B_{\widehat{\xi}}} + \sqrt{B_k}, \qquad \|U\|_{\mathcal{H}_\mathcal{Y}} \leq B_U := 2\sqrt{B_{\widehat{\xi}}} + \sqrt{B_k}, \\ \|U - \mu(V)\|_{\mathcal{H}_\mathcal{Y}} \leq 2B_U.\end{gathered} \tag{29}$$

Step 1: A Bound on $\|\widehat{F}[u]\|^2_{L_2(P_V)}$. Since $k_\mathcal{V}$ is translation invariant, $k_\mathcal{V}(v, v) = \varphi(0) =: \kappa^2$ for every $v \in \mathcal{V}$. We denote by Id the identity operator on $\mathcal{H}_\mathcal{V}$, by C the covariance operator of $k_\mathcal{V}$ and by $\widehat{C}$ its empirical version:

$$C := \mathbb{E}[k_\mathcal{V}(\cdot, V) \otimes k_\mathcal{V}(\cdot, V)], \qquad \widehat{C} := \frac{1}{n}\sum_{i=1}^n k_\mathcal{V}(\cdot, V_{1i}) \otimes k_\mathcal{V}(\cdot, V_{1i}).$$

Since $\mathrm{Tr}(C) = k_\mathcal{V}(v, v) = \kappa^2$, we have $\|C\|_{\mathrm{op}} \leq \kappa^2$. We define the linear maps $S_n : \mathcal{H}_\mathcal{V} \to \mathbb{R}^n$ and $T_u : \mathbb{R}^n \to \mathcal{H}_\mathcal{Y}$ by:

$$S_n(f) := n^{-1/2}(f(V_{11}), \ldots, f(V_{1n})), \quad f \in \mathcal{H}_\mathcal{V}; \qquad T_u(w) := n^{-1/2}\sum_{i=1}^n w_i u_i, \quad w \in \mathbb{R}^n.$$

These maps satisfy:

$$S_n^* S_n = \widehat{C}, \qquad S_n S_n^* = \mathbf{K}_V/n, \qquad \|T_u\|^2_{\mathrm{HS}} = \frac{1}{n}\sum_{i=1}^n \|u_i\|^2_{\mathcal{H}_\mathcal{Y}}.$$

Since $S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1} = (S_nS_n^* + \lambda\mathbf{I}_n)^{-1}S_n$, the definition (27) yields $\widehat{F}[u](v) = A_u k_{\mathcal{V}}(\cdot, v)$ with $A_u \coloneqq T_u S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1}$. Consequently, we have:

$$\begin{aligned}\|\widehat{F}[u]\|^2_{L_2(P_V)} &= \int_{\mathcal{V}} \|A_u k_{\mathcal{V}}(\cdot, v)\|^2_{\mathcal{H}_{\mathcal{Y}}}\, p_V(v)\,\mathrm{d}v = \|A_u C^{1/2}\|^2_{\mathrm{HS}} \\ &\le \|T_u\|^2_{\mathrm{HS}} \|S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1}C^{1/2}\|^2_{\mathrm{op}}.\end{aligned} \tag{30}$$

The spectral theorem yields the following operator norm bounds:

$$\|S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1/2}\|_{\mathrm{op}} \le 1, \quad \|S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1}\|_{\mathrm{op}} \le \frac{1}{2\sqrt{\lambda}}, \quad \|(C+\lambda\,\mathrm{Id})^{-1/2}C^{1/2}\|_{\mathrm{op}} \le 1. \tag{31}$$

Thus, we obtain, for every $u \in \mathcal{H}_{\mathcal{Y}}^n$:

$$\|\widehat{F}[u]\|^2_{L_2(P_V)} \overset{\text{(a)}}{\le} \|T_u\|^2_{\mathrm{HS}} \|S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1}\|^2_{\mathrm{op}} \|C\|_{\mathrm{op}} \overset{\text{(b)}}{\le} \frac{\kappa^2}{4\lambda} \cdot \frac{1}{n}\sum_{i=1}^n \|u_i\|^2_{\mathcal{H}_{\mathcal{Y}}}, \tag{32}$$

where (a) follows from (30) and (b) follows from the formula for $\|T_u\|^2_{\mathrm{HS}}$, the second bound in (31) and $\|C\|_{\mathrm{op}} \le \kappa^2$.

Step 2: The Decomposition of the Error. We denote $\widehat{F}[U] \coloneqq \widehat{F}[u]$ with $u_i = U_i$ and $\widehat{F}[b] \coloneqq \widehat{F}[u]$ with $u_i = b(V_{1i})$. Since $\widehat{\mu}_{\mathrm{RR}} = \widehat{F}[u]$ with $u_i = \widehat{\xi}(Z_{1i}) = U_i + b(V_{1i})$ and $\widehat{F}[u]$ is linear in u, we have $\widehat{\mu}_{\mathrm{RR}} = \widehat{F}[U] + \widehat{F}[b]$, which yields:

$$\|\widehat{\mu}_{\mathrm{RR}} - \mu\|^2_{L_2(P_V)} \le 2\|\widehat{F}[U] - \mu\|^2_{L_2(P_V)} + 2\|\widehat{F}[b]\|^2_{L_2(P_V)}. \tag{33}$$

We denote $R \coloneqq \|\widehat{F}[U] - \mu\|^2_{L_2(P_V)}$. By (32) with $u_i = U_i$, (29) and (28), we have:

$$R \le 2\|\widehat{F}[U]\|^2_{L_2(P_V)} + 2\|\mu\|^2_{L_2(P_V)} \le R_{\max} \coloneqq \frac{\kappa^2 B_U^2}{2\lambda} + 2B_k. \tag{34}$$

We denote $\Delta \coloneqq (C+\lambda\,\mathrm{Id})^{-1/2}(C-\widehat{C})(C+\lambda\,\mathrm{Id})^{-1/2}$ and $E \coloneqq \{\|\Delta\|_{\mathrm{op}} \le 2/3\}$. Since Δ depends only on $V_{11}, \ldots, V_{1n}$, which are independent of $\mathcal{D}_0$, the event E is independent of $\mathcal{D}_0$.

We write T_U for T_u with $u_i = U_i$. We denote:

$$A_\lambda \coloneqq \mathbb{E}[T_U S_n|\mathcal{D}_0]\big(\mathbb{E}[\widehat{C}|\mathcal{D}_0] + \lambda\,\mathrm{Id}\big)^{-1}, \qquad F_\lambda(v) \coloneqq A_\lambda k_{\mathcal{V}}(\cdot, v) \quad \text{for } v \in \mathcal{V}.$$

By the definitions of T_u and S_n, the identity $\mathbb{E}[U|V, \mathcal{D}_0] = \mu(V)$ and the fact that the pairs (V_{1i}, U_i) are copies of (V, U) given $\mathcal{D}_0$, we have:

$$T_U S_n = \frac{1}{n}\sum_{i=1}^n U_i \otimes k_{\mathcal{V}}(\cdot, V_{1i}), \qquad \mathbb{E}[T_U S_n|\mathcal{D}_0] = \mathbb{E}[\mu(V)\otimes k_{\mathcal{V}}(\cdot, V)], \qquad \mathbb{E}[\widehat{C}|\mathcal{D}_0] = C. \tag{35}$$

On E, the triangle inequality in $L_2(P_V)$ yields $R \le 2\|F_\lambda - \mu\|^2_{L_2(P_V)} + 2\|\widehat{F}[U] - F_\lambda\|^2_{L_2(P_V)}$. On E^c, we use (34). By (35), F_λ does not depend on the data. Thus, we obtain:

$$\mathbb{E}[R|\mathcal{D}_0] \le 2\|F_\lambda - \mu\|^2_{L_2(P_V)} + 2\mathbb{E}\Big[\|\widehat{F}[U] - F_\lambda\|^2_{L_2(P_V)}\mathbb{1}_E \,\Big|\, \mathcal{D}_0\Big] + R_{\max}\,\mathbb{P}(E^c|\mathcal{D}_0). \tag{36}$$

We refer to $\|F_\lambda - \mu\|^2_{L_2(P_V)}$ as the *approximation error*, $\mathbb{E}[\|\widehat{F}[U] - F_\lambda\|^2_{L_2(P_V)}\mathbb{1}_E|\mathcal{D}_0]$ as the *statistical error* on E, and $R_{\max}\,\mathbb{P}(E^c|\mathcal{D}_0)$ as the *remainder*. The three terms on the right-hand side of (36), together with $\mathbb{E}[\|\widehat{F}[b]\|^2_{L_2(P_V)}|\mathcal{D}_0]$, are the four terms that we bound in Steps 3–6.

Step 3: The Approximation Error. Let $(e_k)_{k\ge1}$ be an orthonormal basis of $\mathcal{H}_{\mathcal{Y}}$, and denote $c_k(v) \coloneqq \langle \mu(v), e_k\rangle_{\mathcal{H}_{\mathcal{Y}}} = \int_{\mathcal{Y}} e_k(y)\mathfrak{p}^1(y|v)\,\mathrm{d}y$ for $v \in \mathcal{V}$. Lemma 10 with the span of $e_1, \ldots, e_M$ yields $\sum_{k=1}^M \|c_k\|^2_{W^{r,2}(\mathcal{V})} \le B_c$ for every M, with a constant B_c that does not depend on M and is finite by Assumption 6. Letting $M \to \infty$ yields:

$$\sum_{k\ge1} \|c_k\|^2_{W^{r,2}(\mathcal{V})} \le B_c. \tag{37}$$

By (35), the adjoint of $A_\lambda = \mathbb{E}[\mu(V) \otimes k_{\mathcal{V}}(\cdot, V)](C + \lambda \mathrm{Id})^{-1}$ is $A_\lambda^* = (C + \lambda \mathrm{Id})^{-1}\mathbb{E}[k_{\mathcal{V}}(\cdot, V) \otimes \mu(V)]$, which yields, for every k:

$$A_\lambda^* e_k = (C + \lambda \mathrm{Id})^{-1}\mathbb{E}[c_k(V)k_{\mathcal{V}}(\cdot, V)]. \tag{38}$$

According to [FS20, Lemma 14], for the function $A_\lambda^* e_k$ in (38), the bound (37) with $r \leq \tau$ yields, for every k:

$$\mathbb{E}\Big[((A_\lambda^* e_k)(V) - c_k(V))^2\Big] \lesssim \lambda^{r/\tau}\|c_k\|^2_{W^{r,2}(\mathcal{V})}. \tag{39}$$

Since $\langle F_\lambda(v), e_k\rangle_{\mathcal{H}_\mathcal{Y}} = \langle k_{\mathcal{V}}(\cdot, v), A_\lambda^* e_k\rangle_{\mathcal{H}_\mathcal{V}} = (A_\lambda^* e_k)(v)$ for every k and $v \in \mathcal{V}$, it follows from the Parseval identity that:

$$\begin{aligned}
\|F_\lambda - \mu\|^2_{L_2(P_V)} &= \sum_{k\geq 1} \mathbb{E}\Big[(\langle F_\lambda(V), e_k\rangle_{\mathcal{H}_\mathcal{Y}} - c_k(V))^2\Big] \\
&= \sum_{k\geq 1} \mathbb{E}\Big[((A_\lambda^* e_k)(V) - c_k(V))^2\Big] \\
&\overset{(a)}{\lesssim} \lambda^{r/\tau} \sum_{k\geq 1} \|c_k\|^2_{W^{r,2}(\mathcal{V})} \overset{(b)}{\leq} B_c \lambda^{r/\tau},
\end{aligned} \tag{40}$$

where (a) follows from (39) and (b) follows from (37). Since $\lambda \asymp n^{-2\tau/(2r+d_v)}$, we have $\lambda^{r/\tau} \asymp n^{-2r/(2r+d_v)}$. Thus, the approximation error satisfies:

$$\|F_\lambda - \mu\|^2_{L_2(P_V)} \lesssim n^{-2r/(2r+d_v)}. \tag{41}$$

Step 4: The Statistical Error on E. We denote:

$$\zeta_i \coloneqq (U_i - F_\lambda(V_{1i})) \otimes k_{\mathcal{V}}(\cdot, V_{1i}) \quad \text{for } i = 1, \ldots, n, \qquad \Xi \coloneqq \frac{1}{n}\sum_{i=1}^n \zeta_i - \mathbb{E}[\zeta_1|\mathcal{D}_0].$$

We first show that $\|\widehat{F}[U] - F_\lambda\|^2_{L_2(P_V)} \leq 9\|\Xi(C + \lambda \mathrm{Id})^{-1/2}\|^2_{\mathrm{HS}}$ on E. Then we bound $\mathbb{E}[\|\Xi(C + \lambda \mathrm{Id})^{-1/2}\|^2_{\mathrm{HS}}|\mathcal{D}_0]$.

By (35), the identity $F_\lambda(V_{1i}) \otimes k_{\mathcal{V}}(\cdot, V_{1i}) = A_\lambda(k_{\mathcal{V}}(\cdot, V_{1i}) \otimes k_{\mathcal{V}}(\cdot, V_{1i}))$ and the definition of A_λ, we have:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \zeta_i = T_U S_n - A_\lambda \widehat{C}, \\
&\mathbb{E}[\zeta_1|\mathcal{D}_0] = \mathbb{E}[T_U S_n|\mathcal{D}_0] - A_\lambda C = A_\lambda(C + \lambda \mathrm{Id}) - A_\lambda C = \lambda A_\lambda.
\end{aligned}$$

Hence, $\Xi = T_U S_n - A_\lambda(\widehat{C} + \lambda \mathrm{Id})$. We write A_U for A_u with $u_i = U_i$, so that $\widehat{F}[U](v) = A_U k_{\mathcal{V}}(\cdot, v)$ with $A_U = T_U S_n(\widehat{C} + \lambda \mathrm{Id})^{-1}$. Thus, $A_U - A_\lambda = \Xi(\widehat{C} + \lambda \mathrm{Id})^{-1}$ and $\widehat{F}[U](v) - F_\lambda(v) = (A_U - A_\lambda)k_{\mathcal{V}}(\cdot, v)$. On E, the spectrum of the self-adjoint operator $(C + \lambda \mathrm{Id})^{-1/2}(\widehat{C} + \lambda \mathrm{Id})(C + \lambda \mathrm{Id})^{-1/2} = \mathrm{Id} - \Delta$ lies in $[1/3, 5/3]$, which yields:

$$\|(\widehat{C} + \lambda \mathrm{Id})^{-1/2}(C + \lambda \mathrm{Id})^{1/2}\|^2_{\mathrm{op}} = \|(\mathrm{Id} - \Delta)^{-1}\|_{\mathrm{op}} \leq 3. \tag{42}$$

Consequently, we obtain, on E:

$$\begin{aligned}
&\|\widehat{F}[U] - F_\lambda\|^2_{L_2(P_V)} \\
&\overset{(a)}{=} \|\Xi(\widehat{C} + \lambda \mathrm{Id})^{-1} C^{1/2}\|^2_{\mathrm{HS}} \\
&= \|\Xi(C + \lambda \mathrm{Id})^{-1/2} (C + \lambda \mathrm{Id})^{1/2}(\widehat{C} + \lambda \mathrm{Id})^{-1}(C + \lambda \mathrm{Id})^{1/2} (C + \lambda \mathrm{Id})^{-1/2} C^{1/2}\|^2_{\mathrm{HS}} \\
&\overset{(b)}{\leq} \|\Xi(C + \lambda \mathrm{Id})^{-1/2}\|^2_{\mathrm{HS}} \|(\widehat{C} + \lambda \mathrm{Id})^{-1/2}(C + \lambda \mathrm{Id})^{1/2}\|^4_{\mathrm{op}} \|(C + \lambda \mathrm{Id})^{-1/2} C^{1/2}\|^2_{\mathrm{op}} \\
&\overset{(c)}{\leq} 9\|\Xi(C + \lambda \mathrm{Id})^{-1/2}\|^2_{\mathrm{HS}},
\end{aligned} \tag{43}$$

where (a) follows from the equalities in (30) with A_u replaced by $A_U - A_\lambda = \Xi(\widehat{C} + \lambda \mathrm{Id})^{-1}$; (b) follows from $\|(C + \lambda \mathrm{Id})^{1/2}(\widehat{C} + \lambda \mathrm{Id})^{-1}(C + \lambda \mathrm{Id})^{1/2}\|_{\mathrm{op}} = \|(\widehat{C} + \lambda \mathrm{Id})^{-1/2}(C + \lambda \mathrm{Id})^{1/2}\|^2_{\mathrm{op}}$; and (c) follows from (42) and the third bound in (31).

We now bound $\mathbb{E}[\|\Xi(C+\lambda\,\mathrm{Id})^{-1/2}\|_{\mathrm{HS}}^2|\mathcal{D}_0]$. Let $\gamma := \min\{1,(d_v+r)/(2\tau)\}$. Since $\tau \geq r > d_v/2$, γ satisfies $d_v/(2\tau) < \gamma \leq 1$. We use two facts from [FS20, Section 4]:[1]

(i) the i-th largest eigenvalue of C is $O(i^{-2\tau/d_v})$;

(ii) $W^{\gamma\tau,2}(\mathcal{V}) \hookrightarrow L^\infty(\mathcal{V})$.

According to [FS20, Lemmas 11 and 13, Theorem 9], (i) and (ii) yield, respectively, the following bounds for P_V-almost every v, with a constant $\kappa_\gamma > 0$ that depends only on $k_\mathcal{V}$, p_V and γ:

$$\mathrm{Tr}(C(C+\lambda\,\mathrm{Id})^{-1}) \lesssim \lambda^{-d_v/(2\tau)}, \qquad \|(C+\lambda\,\mathrm{Id})^{-1/2}k_\mathcal{V}(\cdot,v)\|_{\mathcal{H}_\mathcal{V}}^2 \leq \kappa_\gamma^2\lambda^{-\gamma}. \tag{44}$$

Since $\mathbb{E}[k_\mathcal{V}(\cdot,V)\otimes k_\mathcal{V}(\cdot,V)] = C$, we have:

$$\begin{aligned}
&\mathbb{E}\|(C+\lambda\,\mathrm{Id})^{-1/2}k_\mathcal{V}(\cdot,V)\|_{\mathcal{H}_\mathcal{V}}^2 \\
&\quad= \mathbb{E}\Big[\mathrm{Tr}\left((C+\lambda\,\mathrm{Id})^{-1/2}(k_\mathcal{V}(\cdot,V)\otimes k_\mathcal{V}(\cdot,V))(C+\lambda\,\mathrm{Id})^{-1/2}\right)\Big] \\
&\quad= \mathrm{Tr}(C(C+\lambda\,\mathrm{Id})^{-1}).
\end{aligned} \tag{45}$$

Thus, we obtain:

$$\begin{aligned}
\mathbb{E}\Big[\|\Xi(C+\lambda\,\mathrm{Id})^{-1/2}\|_{\mathrm{HS}}^2\,\Big|\,\mathcal{D}_0\Big] &\overset{(a)}{\leq} \frac{1}{n}\mathbb{E}\Big[\|\zeta_1(C+\lambda\,\mathrm{Id})^{-1/2}\|_{\mathrm{HS}}^2\,\Big|\,\mathcal{D}_0\Big] \\
&\overset{(b)}{=} \frac{1}{n}\mathbb{E}\Big[\|U-F_\lambda(V)\|_{\mathcal{H}_\mathcal{Y}}^2\|(C+\lambda\,\mathrm{Id})^{-1/2}k_\mathcal{V}(\cdot,V)\|_{\mathcal{H}_\mathcal{V}}^2\,\Big|\,\mathcal{D}_0\Big] \\
&\overset{(c)}{\leq} \frac{2}{n}\mathbb{E}\Big[\|U-\mu(V)\|_{\mathcal{H}_\mathcal{Y}}^2\|(C+\lambda\,\mathrm{Id})^{-1/2}k_\mathcal{V}(\cdot,V)\|_{\mathcal{H}_\mathcal{V}}^2\,\Big|\,\mathcal{D}_0\Big] \\
&\quad+\frac{2}{n}\mathbb{E}\Big[\|\mu(V)-F_\lambda(V)\|_{\mathcal{H}_\mathcal{Y}}^2\|(C+\lambda\,\mathrm{Id})^{-1/2}k_\mathcal{V}(\cdot,V)\|_{\mathcal{H}_\mathcal{V}}^2\Big] \\
&\overset{(d)}{\leq} \frac{8B_U^2\,\mathrm{Tr}(C(C+\lambda\,\mathrm{Id})^{-1})+2\kappa_\gamma^2\lambda^{-\gamma}\|F_\lambda-\mu\|_{L_2(P_V)}^2}{n},
\end{aligned} \tag{46}$$

where (a) holds because, given $\mathcal{D}_0$, $\Xi(C+\lambda\,\mathrm{Id})^{-1/2} = \frac{1}{n}\sum_{i=1}^n \zeta_i(C+\lambda\,\mathrm{Id})^{-1/2} - \mathbb{E}[\zeta_1(C+\lambda\,\mathrm{Id})^{-1/2}|\mathcal{D}_0]$ is the centered average of the i.i.d. Hilbert-Schmidt operators $\zeta_i(C+\lambda\,\mathrm{Id})^{-1/2}$; (b) follows from $\zeta_1(C+\lambda\,\mathrm{Id})^{-1/2} = (U_1 - F_\lambda(V_{11}))\otimes(C+\lambda\,\mathrm{Id})^{-1/2}k_\mathcal{V}(\cdot,V_{11})$, the identity $\|a\otimes b\|_{\mathrm{HS}} = \|a\|\|b\|$ and the fact that (V_{11},U_1) is a copy of (V,U) given $\mathcal{D}_0$; (c) follows from $\|U-F_\lambda(V)\|_{\mathcal{H}_\mathcal{Y}}^2 \leq 2\|U-\mu(V)\|_{\mathcal{H}_\mathcal{Y}}^2 + 2\|\mu(V)-F_\lambda(V)\|_{\mathcal{H}_\mathcal{Y}}^2$; and (d) follows from the third bound in (29), the second bound in (44) and (45). Since $\lambda \asymp n^{-2\tau/(2r+d_v)}$ and $2\gamma\tau \leq d_v + r$, we have:

$$n\lambda^\gamma \gtrsim n^{r/(2r+d_v)} \geq 1. \tag{47}$$

Thus, (43), (46), the first bound in (44) and (40) yield:

$$\begin{aligned}
\mathbb{E}\Big[\|\widehat{F}[U]-F_\lambda\|_{L_2(P_V)}^2\mathbb{1}_E\,\Big|\,\mathcal{D}_0\Big] &\leq 9\,\mathbb{E}\Big[\|\Xi(C+\lambda\,\mathrm{Id})^{-1/2}\|_{\mathrm{HS}}^2\,\Big|\,\mathcal{D}_0\Big] \\
&\lesssim \frac{\lambda^{-d_v/(2\tau)}}{n}+\frac{\lambda^{r/\tau}}{n\lambda^\gamma} \lesssim n^{-2r/(2r+d_v)}.
\end{aligned} \tag{48}$$

Step 5: The Remainder and the Bound on $\mathbb{E}[R|\mathcal{D}_0]$. According to [FS20, Lemma 17], the conditions that Δ depends only on the independent copies $V_{11},\ldots,V_{1n}$ of V and that the embedding in (ii) holds yield the following bound for every $s \geq 1$, with probability at least $1-2e^{-s}$:

$$\|\Delta\|_{\mathrm{op}} \leq \frac{4\kappa_\gamma^2 s\,g_\lambda}{3n\lambda^\gamma} + \sqrt{\frac{2\kappa_\gamma^2 s\,g_\lambda}{n\lambda^\gamma}}, \tag{49}$$

[1] In [FS20, Section 4], the facts hold for domains with C^∞ boundaries and densities that are bounded above and below. Here, p_V is bounded above and below by Assumption 4 and the assumption $\inf_{v\in\mathcal{V}} p_V(v) > 0$ of the theorem. We can extend these facts to the Lipschitz boundary of $\mathcal{V}$ via Stein's extension operator, which maps $W^{k,2}(\mathcal{V})$ continuously into $W^{k,2}(\mathbb{R}^{d_v})$ for every integer $k \geq 0$ [Ste70, Chapter VI, Theorem 5]. We then extend this operator to every real order $t > 0$ by the standard interpolation argument between the integer orders (see, e.g., [CWHM15, Lemma 4.2]). Thus, we can extend the facts to Sobolev spaces with fractional orders.

where κ_γ is the constant in (44) and $g_\lambda := \log\big(2e\ \mathrm{Tr}(C(C+\lambda\,\mathrm{Id})^{-1})(\|C\|_{\mathrm{op}}+\lambda)/\|C\|_{\mathrm{op}}\big)$. Let $n_1 \in \mathbb{N}$ be sufficiently large so that $\lambda \asymp n^{-2\tau/(2r+d_v)} \le \min\{1, \|C\|_{\mathrm{op}}\}$ for every $n \ge n_1$. Then, for such n, we have:

$$g_\lambda \le \log\big(4e\ \mathrm{Tr}(C(C+\lambda\,\mathrm{Id})^{-1})\big) \le \log(4e\ \mathrm{Tr}(C)/\lambda) = \log(4e\kappa^2/\lambda).$$

Let $\nu := r/(2(2r+d_v))$. By (47) and $\lambda \asymp n^{-2\tau/(2r+d_v)}$, we have $n\lambda^\gamma \gtrsim n^{2\nu}$ and $\log(4e\kappa^2/\lambda) \lesssim \log n$. Consequently, there is $n_2 \ge n_1$ such that $8\kappa_\gamma^2 n^\nu \log(4e\kappa^2/\lambda) \le n\lambda^\gamma$ for every $n \ge n_2$. For such n and $s = n^\nu$, we have $\kappa_\gamma^2 s\, g_\lambda/(n\lambda^\gamma) \le 1/8$. Thus, the right-hand side of (49) is at most $2/3$, which yields $\mathbb{P}(E) \ge 1 - 2e^{-n^\nu}$. Moreover, $\lambda \le 1$ yields $R_{\max} \lesssim \lambda^{-1} \asymp n^{2\tau/(2r+d_v)}$, while e^{-n^ν} decays faster than any power of n. Since E is independent of $\mathcal{D}_0$, we obtain, for every $n \ge n_2$:

$$\mathbb{P}(E^c|\mathcal{D}_0) = \mathbb{P}(E^c) \le 2e^{-n^\nu}, \qquad R_{\max}\,\mathbb{P}(E^c|\mathcal{D}_0) \lesssim \lambda^{-1}e^{-n^\nu} \lesssim n^{-2r/(2r+d_v)}. \tag{50}$$

Combining (36), (41), (48) and (50), we obtain, for every $n \ge n_2$, with a constant that does not depend on $\mathcal{D}_0$:

$$\mathbb{E}[R|\mathcal{D}_0] = \mathbb{E}\Big[\|\widehat{F}[U] - \mu\|^2_{L_2(P_V)}\,\Big|\,\mathcal{D}_0\Big] \lesssim n^{-2r/(2r+d_v)}. \tag{51}$$

And for all $n < n_2$, we use $R \le R_{\max}$. Thus, (51) holds for every $n \ge 1$.

Step 6: The Bound on $\mathbb{E}[\|\widehat{F}[b]\|^2_{L_2(P_V)}|\mathcal{D}_0]$. On E, we have, for every $u \in \mathcal{H}^n_{\mathcal{Y}}$:

$$\begin{aligned}
&\|\widehat{F}[u]\|^2_{L_2(P_V)} \\
&\overset{(a)}{\le} \|T_u\|^2_{\mathrm{HS}}\|S_n(\widehat{C}+\lambda\,\mathrm{Id})^{-1/2}\|^2_{\mathrm{op}}\|(\widehat{C}+\lambda\,\mathrm{Id})^{-1/2}(C+\lambda\,\mathrm{Id})^{1/2}\|^2_{\mathrm{op}}\|(C+\lambda\,\mathrm{Id})^{-1/2}C^{1/2}\|^2_{\mathrm{op}} \\
&\overset{(b)}{\le} \frac{3}{n}\sum_{i=1}^n \|u_i\|^2_{\mathcal{H}_{\mathcal{Y}}},
\end{aligned} \tag{52}$$

where (a) follows from (30) and (b) follows from the formula for $\|T_u\|^2_{\mathrm{HS}}$, the first and third bounds in (31) and (42). Recall that $b = \widetilde{\mu}_{Y^1|V} - \mu$ depends only on $\mathcal{D}_0$ and satisfies $\sup_{v\in\mathcal{V}}\|b(v)\|_{\mathcal{H}_{\mathcal{Y}}} \le \sqrt{B_{\widehat{\xi}}} + \sqrt{B_k}$. Since E is independent of $\mathcal{D}_0$, (52) on E, (32) on E^c, both with $u_i = b(V_{1i})$, and (50) yield, for every $n \ge n_2$:

$$\begin{aligned}
\mathbb{E}\Big[\|\widehat{F}[b]\|^2_{L_2(P_V)}\,\Big|\,\mathcal{D}_0\Big] &\le 3\,\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\|b(V_{1i})\|^2_{\mathcal{H}_{\mathcal{Y}}}\,\middle|\,\mathcal{D}_0\right] + \frac{\kappa^2(\sqrt{B_{\widehat{\xi}}}+\sqrt{B_k})^2}{4\lambda}\,\mathbb{P}(E^c) \\
&\le 3\|b\|^2_{L_2(P_V)} + \frac{\kappa^2(\sqrt{B_{\widehat{\xi}}}+\sqrt{B_k})^2}{2\lambda}\,e^{-n^\nu} \\
&\lesssim \|b\|^2_{L_2(P_V)} + n^{-2r/(2r+d_v)} \\
&= \|\widetilde{\mu}_{Y^1|V} - \mu\|^2_{L_2(P_V)} + n^{-2r/(2r+d_v)}.
\end{aligned} \tag{53}$$

And for all $n < n_2$, we use (32). Thus, (53) holds for every $n \ge 1$.

Conclusion. Finally, taking the expectation of (33) and applying the bounds (51) and (53) yield:

$$\mathbb{E}\big[\|\widehat{\mu}_{\mathrm{RR}}(V) - \mu(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}\big] \lesssim n^{-\frac{2r}{2r+d_v}} + \mathbb{E}\big[\|\widetilde{\mu}_{Y^1|V}(V) - \mu(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}\big].$$

Since the assumptions of the theorem include $\pi(x), \widehat{\pi}(x) \in [\varepsilon, 1)$ and $\sup_{x\in\mathcal{X}}\|\widehat{\mu}_0(x)\|_{\mathcal{H}_{\mathcal{Y}}} \le B_{\mu_0}$, we can use Lemma 1 to obtain $\mathbb{E}[\|\widetilde{\mu}_{Y^1|V}(V) - \mu(V)\|^2_{\mathcal{H}_{\mathcal{Y}}}] \lesssim \min\big\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\big\}$, which completes the proof. ■

L Tools for Generalization Bounds

Notations. Recall the two datasets $\mathcal{D}_0 = \{Z_{0i}\}_{i=1}^n$ and $\mathcal{D}_1 = \{Z_{1i}\}_{i=1}^n$, where $Z_{0i} = (X_{0i}, A_{0i}, Y_{0i})$ and $Z_{1i} = (X_{1i}, A_{1i}, Y_{1i})$ are independent draws from P. Let $V_{1i} = \eta(X_{1i})$.

Let $Z = (X, A, Y) \sim P$ independent of $\mathcal{D}_0$ and $\mathcal{D}_1$, and let $V = \eta(X)$. Recall from Appendix K.1 that P_V denotes the distribution of V and that $\|h\|^2_{L_2(P_V)} = \int_{\mathcal{V}} \|h(v)\|^2_{\mathcal{H}_\mathcal{Y}} \, p_V(v) \, \mathrm{d}v$ for $h \in (\mathcal{V} \to \mathcal{H}_\mathcal{Y})$. The variable Z takes values in $\mathcal{Z} \coloneqq \mathcal{X} \times \{0,1\} \times \mathcal{Y}$, whose elements we denote by $z = (x, a, y)$. For a bounded function $g \colon \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$, which may depend on $\mathcal{D}_0$ and $\mathcal{D}_1$, we denote $\mathbb{E}_n[g(V,Z)] \coloneqq \frac{1}{n}\sum_{i=1}^n g(V_{1i}, Z_{1i})$. The expectation of $g(V,Z)$ over Z is $\mathbb{E}[g(V,Z) \mid \mathcal{D}_0, \mathcal{D}_1]$. The pseudo-outcome $\widehat{\xi}$ is the map defined in (3), where, by the assumptions of Theorem 1, the first-stage estimator satisfies $\sup_{x \in \mathcal{X}} \|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$. We assume throughout that the kernel satisfies $\sup_{y \in \mathcal{Y}} k_\mathcal{Y}(y,y) \leq B_k$ for some constant $B_k > 0$. For a map f with values in $\mathbb{R}^M$, we denote by $f_k(v)$ the k-th component of $f(v)$. When the map itself has a subscript c, we denote by $f_{c,k}(v)$ the k-th component of $f_c(v)$. We denote the cardinality of a finite set by $|\cdot|$.

We start with some standard tools that are useful for obtaining excess risk bounds for classes of neural networks.

Definition 11 (Metric Entropy)**.** *Let $\mathcal{S}$ be a set. Let $\mathcal{F} \subseteq (\mathcal{S} \to \mathbb{R}^M)$ for some $M \geq 1$, and let Q be a probability measure on $\mathcal{S}$. For any $\varepsilon > 0$ and $p \in [1, \infty)$, an* (ε, p)-cover *of $\mathcal{F}$ in $L_p(Q)$ is a set $\mathcal{F}' \subseteq (\mathcal{S} \to \mathbb{R}^M)$ such that for any $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}'$ with*

$$\left(\int_\mathcal{S} \|f(s) - f'(s)\|^p \, \mathrm{d}Q(s) \right)^{1/p} \leq \varepsilon.$$

The (ε, p)-covering number*, denoted by $\mathcal{N}_p(\mathcal{F}, \varepsilon, Q)$, is the cardinality of the smallest such (ε, p)-cover. The* (ε, p)-metric entropy *is the logarithm of the covering number:*

$$\mathcal{H}_p(\mathcal{F}, \varepsilon, Q) \coloneqq \log \mathcal{N}_p(\mathcal{F}, \varepsilon, Q).$$

We consider a class $\mathcal{F}^M \subseteq (\mathcal{V} \to \mathbb{R}^M)$ and a linear isometry $\iota : \mathbb{R}^M \to \mathcal{H}_\mathcal{Y}$. Let $\ell(h, z) \coloneqq \|h - \widehat{\xi}(z)\|^2_{\mathcal{H}_\mathcal{Y}}$, for $h \in \mathcal{H}_\mathcal{Y}$ and $z \in \mathcal{Z}$, be the squared loss. Denote $\mathcal{L}_\mu(v, z) \coloneqq \ell(\mu(v), z)$ for every map μ with values in $\mathcal{H}_\mathcal{Y}$. Let $\widehat{f}$ be an empirical risk minimizer and $f^\star$ be a minimizer of the population risk, i.e.,

$$\widehat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{F}^M} \mathbb{E}_n[\mathcal{L}_{\iota \circ f}(V, Z)], \qquad f^\star \coloneqq \operatorname*{arg\,min}_{f \in \mathcal{F}^M} \mathbb{E}[\mathcal{L}_{\iota \circ f}(V, Z) \mid \mathcal{D}_0]. \tag{54}$$

We shall assume that such minimizers exist.

We now bound the metric entropy of a class of neural networks. For widths $\mathbf{d} = (d_0, \ldots, d_L)$ with $d_0 = d_v$ and a constant $B \in (0, \infty]$, we denote by $\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B)$ the class of the fully connected networks (12) with outputs bounded by B:

$$\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B) \coloneqq \left\{ \psi_\theta : \theta \in \textstyle\prod_{i=1}^L \left(\mathbb{R}^{d_i \times d_{i-1}} \times \mathbb{R}^{d_i} \right), \, \sup_{v \in \mathcal{V}} \|\psi_\theta(v)\| \leq B \right\}. \tag{55}$$

A common approach to bounding the metric entropy treats a network with M outputs as M scalar-valued functions. However, this approach yields a bound on the metric entropy that is linear in M. To reduce this dependence on M, we instead view these networks as functions that take the output index as an input. The following lemma provides the resulting bound:

Lemma 3 (Metric Entropy of Fully Connected Networks with M Outputs)**.** *Let $\mathbf{d} \coloneqq (d_0, \ldots, d_L)$ with $d_0 = d_v$, $d_L = M$ and $W \geq 2$, and let $B > 0$. Then there is a universal constant $c_{\mathrm{FC}} \geq 1$ such that, for every probability measure Q on $\mathcal{V}$ and every $\varepsilon \in (0, B]$,*

$$\mathcal{H}_2(\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B), \varepsilon, Q) \leq c_{\mathrm{FC}} \, W L \log W \log \left(\frac{4B\sqrt{M}}{\varepsilon} \right).$$

Proof. We denote by $\mathbb{1}_k \in \mathbb{R}^M$ the k-th one-hot vector. For $\psi \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B)$, we denote by $\psi_k(v)$ the k-th coordinate of $\psi(v) \in \mathbb{R}^M$. We define $g_\psi \in (\mathcal{V} \times \{\mathbb{1}_1, \ldots, \mathbb{1}_M\} \to \mathbb{R})$ by $g_\psi(v, \mathbb{1}_k) \coloneqq \psi_k(v)$.

We first show that a ReLU network with $O(W)$ parameters computes g_ψ. We denote by $h(v)$ the output of the last hidden layer, so that $\psi(v) = \mathbf{W}_L h(v) + \mathfrak{b}_L$. Moreover, we denote by $w_k^\top$ the k-th row of $\mathbf{W}_L$. The network that we construct has the input $(v, u) \in \mathbb{R}^{d_v + M}$ and consists of the following three components:

1. the hidden layers of ψ, which take v as input and compute $h(v)$;
2. a layer of $2M$ ReLU units $s_k^{\pm} := \sigma_{\mathrm{relu}}(\pm(w_k^\top h(v) + (\mathfrak{b}_L)_k) + 2Bu_k - 2B)$, $k = 1, \dots, M$, where u_k is the k-th component of u;
3. an output unit with value $\sum_{k=1}^{M}(s_k^+ - s_k^-)$.

At $u = \mathbb{1}_k$, the network above computes $g_\psi(v, \mathbb{1}_k)$. Indeed, since $|\psi_j(v)| \le \|\psi(v)\| \le B$ for every j, we have:

$$s_j^{\pm} = \sigma_{\mathrm{relu}}(\pm\psi_j(v) - 2B) = 0 \quad \text{for } j \neq k,$$
$$s_k^+ - s_k^- = \sigma_{\mathrm{relu}}(\psi_k(v)) - \sigma_{\mathrm{relu}}(-\psi_k(v)) = \psi_k(v).$$

Components 2 and 3 have $2M(d_{L-1} + 3) \lesssim M(d_{L-1} + 1)$ parameters. Since the output layer of ψ has $M(d_{L-1} + 1) \le W$ parameters, this network has $\lesssim W$ parameters and $L + 1$ layers, with which [BHLM19, Theorem 7] applies and yields that the class of functions on $\mathbb{R}^{d_v+M}$ computed by these networks has VC-dimension at most $d_{\mathrm{VC}} := c\,WL\log W$ for a universal constant $c \ge 1$. From [AB99, Theorem 14.1] and its subsequent remark, d_{VC} is also an upper bound on the pseudo-dimension of this class. Consequently, the class $\mathcal{G}_{\mathrm{FC}} := \{g_\psi : \psi \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B)\}$ has pseudo-dimension at most d_{VC}.

Next, we convert this bound into a bound on the metric entropy of $\mathcal{G}_{\mathrm{FC}}$. We use the covering numbers of Definition 11 for real-valued functions on $\mathcal{S} = \mathcal{V} \times \{\mathbb{1}_1, \dots, \mathbb{1}_M\}$ and the product $Q \otimes \mathrm{Unif}\{\mathbb{1}_1, \dots, \mathbb{1}_M\}$ of Q and the uniform distribution on $\{\mathbb{1}_1, \dots, \mathbb{1}_M\}$. Since $|g| \le B$ for every $g \in \mathcal{G}_{\mathrm{FC}}$, the class $\{(g + B)/(2B) : g \in \mathcal{G}_{\mathrm{FC}}\}$ takes values in $[0, 1]$ and has pseudo-dimension at most d_{VC}. By [Hau95, Corollary 3], for every $\varepsilon' \in (0, B]$, this class has an $((\varepsilon'/(2B))^2, 1)$-cover in L_1 of this product measure with at most $e(d_{\mathrm{VC}} + 1)(8eB^2/\varepsilon'^2)^{d_{\mathrm{VC}}}$ elements, all of which belong to the class. This cover is also an $(\varepsilon'/(2B), 2)$-cover in L_2 of this product measure, since its elements take values in $[0, 1]$. This yields:

$$\log \mathcal{N}_2(\mathcal{G}_{\mathrm{FC}}, \varepsilon', Q \otimes \mathrm{Unif}\{\mathbb{1}_1, \dots, \mathbb{1}_M\}) \le 2d_{\mathrm{VC}} \log(4eB/\varepsilon'), \qquad 0 < \varepsilon' \le B.$$

Finally, for $\psi, \psi' \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B)$, we have the following identity:

$$\int_{\mathcal{V}} \|\psi(v) - \psi'(v)\|^2 \,\mathrm{d}Q(v) = M \cdot \frac{1}{M} \sum_{k=1}^{M} \int_{\mathcal{V}} (g_\psi(v, \mathbb{1}_k) - g_{\psi'}(v, \mathbb{1}_k))^2 \,\mathrm{d}Q(v).$$

Taking $\varepsilon' = \varepsilon/\sqrt{M}$ yields the claim with $c_{\mathrm{FC}} = 2(1 + 1/\log 4)c$. □

M Proof of the Upper Bound for the Deep Feature Estimator (Theorem 3)

M.1 Notations and Setup

We begin by clarifying notation. Recall from Section 5 that the Deep Feature estimator uses a neural network $\psi_\theta : \mathcal{V} \to \mathbb{R}^M$ parameterized by $\theta \in \Theta$. For the theoretical analysis, we consider the class $\mathcal{F}^M := \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_{\mathcal{F}}) = \{\psi_\theta : \theta \in \Theta\}$ as defined in (55), with $\mathbf{d} = (d_v, d_1, \dots, d_{L-1}, M)$ and $\Theta = \{\theta : \sup_{v \in \mathcal{V}} \|\psi_\theta(v)\| \le B_{\mathcal{F}}\}$, where $B_{\mathcal{F}} > 0$ is a specified bound on the network output to prevent overfitting.

Recall from Appendix L that $\ell(h, z) = \|h - \widehat{\xi}(z)\|^2_{\mathcal{H}_\mathcal{Y}}$ is the squared loss in the RKHS $\mathcal{H}_\mathcal{Y}$. The estimator of Theorem 3 is $\widehat{\mu}_{\mathrm{DF}} = \widehat{C}\psi_{\widehat{\theta}}$ with $\widehat{C} := \widehat{C}_{\psi_{\widehat{\theta}}}$ as defined in (7), where $\psi_\theta \in \mathcal{F}^M$. We denote by $J_n(\theta, C)$ the objective of (6):

$$J_n(\theta, C) := \frac{1}{n} \sum_{i=1}^{n} \|C\psi_\theta(V_{1i}) - \widehat{\xi}(Z_{1i})\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda_1 \|C\|^2_{\mathrm{HS}}.$$

Hence, by our definitions of $\widehat{\theta}$ and $\widehat{C}$, we have $J_n(\widehat{\theta}, \widehat{C}) = \min_{\theta, C} J_n(\theta, C)$.

M.2 Preliminary Lemma

The following lemma provides an oracle inequality for the Deep Feature estimator. In this lemma and in the proof of Theorem 3, we denote $\lambda := \lambda_1$ and $\widetilde{\mu}_{Y^1|V}(v) := \mathbb{E}[\widehat{\xi}(Z)|V = v, \mathcal{D}_0]$.

Lemma 4 (Oracle Inequality for the Deep Feature Estimator)**.** *Let the assumptions of Theorem 3 hold with $\lambda \leq 1$ and $n \geq 2$, and let $B_{\widehat{\xi}}$ be as in Lemma 2. Then, there exists a constant $c_{\mathrm{DF}} > 0$, which depends only on $B_{\widehat{\xi}}$ and $B_{\mathcal{F}}$, such that, for every $\theta_c \in \Theta$ and every $C_c \in \mathcal{BL}(\mathbb{R}^M; \mathcal{H}_\mathcal{Y})$ that do not depend on $\mathcal{D}_1$, the estimator $\widehat{\mu}_{\mathrm{DF}}$ satisfies:*

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{DF}} - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} \mid \mathcal{D}_0] \leq \|C_c\psi_{\theta_c} - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} + \lambda\|C_c\|^2_{\mathrm{HS}} + \frac{c_{\mathrm{DF}}}{\lambda}\left(\frac{WL\log W\log(Mn)}{n}\right)^{1/2}.$$

Proof. By Lemma 2, $\|\widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \leq \sqrt{B_{\widehat{\xi}}}$ for all z. We first bound the Hilbert-Schmidt norm of $\widehat{C}$. Since $\widehat{C}$ minimizes $J_n(\widehat{\theta}, \cdot)$ over $\mathcal{BL}(\mathbb{R}^M; \mathcal{H}_\mathcal{Y})$, comparing $\widehat{C}$ with $C = 0$ yields:

$$\lambda\|\widehat{C}\|^2_{\mathrm{HS}} \leq J_n(\widehat{\theta}, \widehat{C}) \leq J_n(\widehat{\theta}, 0) = \frac{1}{n}\sum_{i=1}^n \|\widehat{\xi}(Z_{1i})\|^2_{\mathcal{H}_\mathcal{Y}} \leq B_{\widehat{\xi}}.$$

Consequently, $\|\widehat{C}\|_{\mathrm{HS}} \leq R_\lambda := \sqrt{B_{\widehat{\xi}}/\lambda}$. Let $\mathcal{F}_{\mathrm{DF}}$ be the following function class:

$$\mathcal{F}_{\mathrm{DF}} := \left\{C\psi : \psi \in \mathcal{F}^M,\ C \in \mathcal{BL}(\mathbb{R}^M; \mathcal{H}_\mathcal{Y}),\ \|C\|_{\mathrm{HS}} \leq R_\lambda\right\}.$$

Thus, $\widehat{\mu}_{\mathrm{DF}} \in \mathcal{F}_{\mathrm{DF}}$ and $\|\widehat{\mu}_{\mathrm{DF}}(v)\|_{\mathcal{H}_\mathcal{Y}} \leq R_\lambda B_\mathcal{F}$ for all $v \in \mathcal{V}$.

Next, we derive an oracle inequality. For the θ_c and C_c in the statement of the lemma, we denote $\psi_c := \psi_{\theta_c}$ and $f_c := C_c\psi_c$. Recall from Appendix L that $\mathbb{E}_n[g(V, Z)]$ is the average of $g(V_{1i}, Z_{1i})$ over the pairs of $\mathcal{D}_1$. Since $J_n(\widehat{\theta}, \widehat{C}) \leq J_n(\theta_c, C_c)$, we have:

$$\mathbb{E}_n[\mathcal{L}_{\widehat{\mu}_{\mathrm{DF}}}(V, Z)] \leq \mathbb{E}_n[\mathcal{L}_{f_c}(V, Z)] + \lambda\|C_c\|^2_{\mathrm{HS}}. \tag{56}$$

Recall also from Appendix L that Z is independent of $\mathcal{D}_0$ and $\mathcal{D}_1$. Thus, we have $\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0, \mathcal{D}_1] = \mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] = \widetilde{\mu}_{Y^1|V}(V)$. Consequently, expanding the square around $\widetilde{\mu}_{Y^1|V}(V)$ yields the identity $\mathbb{E}[\mathcal{L}_f(V, Z) \mid \mathcal{D}_0, \mathcal{D}_1] = \|f - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} + \mathbb{E}[\|\widehat{\xi}(Z) - \widetilde{\mu}_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}} \mid \mathcal{D}_0]$ for every bounded map $f \in (\mathcal{V} \to \mathcal{H}_\mathcal{Y})$, which may depend on $\mathcal{D}_0$ and $\mathcal{D}_1$. With $f = \widehat{\mu}_{\mathrm{DF}}$ and $f = f_c$, we obtain:

$$\mathbb{E}[\mathcal{L}_{\widehat{\mu}_{\mathrm{DF}}}(V, Z) \mid \mathcal{D}_0, \mathcal{D}_1] = \|\widehat{\mu}_{\mathrm{DF}} - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} + \mathbb{E}[\|\widehat{\xi}(Z) - \widetilde{\mu}_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}} \mid \mathcal{D}_0], \tag{57}$$

$$\mathbb{E}[\mathcal{L}_{f_c}(V, Z) \mid \mathcal{D}_0] = \|f_c - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} + \mathbb{E}[\|\widehat{\xi}(Z) - \widetilde{\mu}_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}} \mid \mathcal{D}_0]. \tag{58}$$

Combining (56), (57) and (58) yields:

$$\begin{aligned}\|\widehat{\mu}_{\mathrm{DF}} - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} - \|f_c - \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} \\ \leq \mathbb{E}[\mathcal{L}_{\widehat{\mu}_{\mathrm{DF}}}(V, Z) \mid \mathcal{D}_0, \mathcal{D}_1] - \mathbb{E}_n[\mathcal{L}_{\widehat{\mu}_{\mathrm{DF}}}(V, Z)] \\ + \mathbb{E}_n[\mathcal{L}_{f_c}(V, Z)] - \mathbb{E}[\mathcal{L}_{f_c}(V, Z) \mid \mathcal{D}_0] + \lambda\|C_c\|^2_{\mathrm{HS}}.\end{aligned} \tag{59}$$

We now take the conditional expectation given $\mathcal{D}_0$. Since f_c does not depend on $\mathcal{D}_1$, we have $\mathbb{E}\left[\mathbb{E}_n[\mathcal{L}_{f_c}(V, Z)] \,\middle|\, \mathcal{D}_0\right] = \mathbb{E}[\mathcal{L}_{f_c}(V, Z) \mid \mathcal{D}_0]$. To treat the difference of the first two terms on the right-hand side of (59), we define $\bar{\mathcal{L}}_f$ for $f \in (\mathcal{V} \to \mathcal{H}_\mathcal{Y})$ by:

$$\bar{\mathcal{L}}_f(v, z) := \|f(v)\|^2_{\mathcal{H}_\mathcal{Y}} - 2\langle f(v), \widehat{\xi}(z)\rangle_{\mathcal{H}_\mathcal{Y}} = \mathcal{L}_f(v, z) - \|\widehat{\xi}(z)\|^2_{\mathcal{H}_\mathcal{Y}}.$$

By this identity, and since $\|\widehat{\xi}(z)\|^2_{\mathcal{H}_\mathcal{Y}}$ does not depend on $\mathcal{D}_1$, the conditional expectation of this difference given $\mathcal{D}_0$ does not change when we replace $\mathcal{L}_{\widehat{\mu}_{\mathrm{DF}}}$ by $\bar{\mathcal{L}}_{\widehat{\mu}_{\mathrm{DF}}}$.

We now bound the conditional expectation of $\mathbb{E}[\bar{\mathcal{L}}_{\widehat{\mu}_{\mathrm{DF}}}(V,Z) \mid \mathcal{D}_0, \mathcal{D}_1] - \mathbb{E}_n[\bar{\mathcal{L}}_{\widehat{\mu}_{\mathrm{DF}}}(V,Z)]$ given $\mathcal{D}_0$. Since $\widehat{\mu}_{\mathrm{DF}} \in \mathcal{F}_{\mathrm{DF}}$, we have:

$$\begin{aligned}&\mathbb{E}\big[\mathbb{E}[\bar{\mathcal{L}}_{\widehat{\mu}_{\mathrm{DF}}}(V,Z) \mid \mathcal{D}_0, \mathcal{D}_1] - \mathbb{E}_n[\bar{\mathcal{L}}_{\widehat{\mu}_{\mathrm{DF}}}(V,Z)] \,\big|\, \mathcal{D}_0\big] \\ &\quad\le \mathbb{E}\left[\sup_{f \in \mathcal{F}_{\mathrm{DF}}} \big|\mathbb{E}[\bar{\mathcal{L}}_f(V,Z) \mid \mathcal{D}_0] - \mathbb{E}_n[\bar{\mathcal{L}}_f(V,Z)]\big| \,\middle|\, \mathcal{D}_0\right].\end{aligned} \tag{60}$$

On $\mathcal{F}_{\mathrm{DF}}$, we have $|\bar{\mathcal{L}}_f| \le R_\lambda^2 B_{\mathcal{F}}^2 + 2R_\lambda B_{\mathcal{F}} \sqrt{B_{\widehat{\xi}}}$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher variables, and denote

$$\mathrm{Rad}_n := \mathbb{E}_\varepsilon\left[\sup_{f \in \mathcal{F}_{\mathrm{DF}}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, \bar{\mathcal{L}}_f(V_{1i}, Z_{1i})\right|\right].$$

By the symmetrization inequality [Wai19, Chapter 4], the condition that $\bar{\mathcal{L}}_f$ is bounded on $\mathcal{F}_{\mathrm{DF}}$ yields:

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}_{\mathrm{DF}}} \big|\mathbb{E}[\bar{\mathcal{L}}_f(V,Z) \mid \mathcal{D}_0] - \mathbb{E}_n[\bar{\mathcal{L}}_f(V,Z)]\big| \,\middle|\, \mathcal{D}_0\right] \le 2\,\mathbb{E}[\mathrm{Rad}_n \mid \mathcal{D}_0]. \tag{61}$$

Taking the conditional expectation of (59) and inserting (60) and (61) yield:

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{DF}} - \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 \mid \mathcal{D}_0] \le \|f_c - \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 + \lambda \|C_c\|_{\mathrm{HS}}^2 + 2\,\mathbb{E}[\mathrm{Rad}_n \mid \mathcal{D}_0]. \tag{62}$$

It remains to bound Rad_n, which will be done via a standard covering argument combined with a maximal inequality. For $\psi \in \mathcal{F}^M$, we write $\psi_i := \psi(V_{1i}) \in \mathbb{R}^M$ and $\widehat{\xi}_i := \widehat{\xi}(Z_{1i})$. Since $\langle C\psi_i, \widehat{\xi}_i\rangle_{\mathcal{H}_\mathcal{Y}} = \langle C, \widehat{\xi}_i \otimes \psi_i\rangle_{\mathrm{HS}}$ and $\|C\psi_i\|_{\mathcal{H}_\mathcal{Y}}^2 = \psi_i^\top C^* C\, \psi_i$, we have:

$$\bar{\mathcal{L}}_{C\psi}(V_{1i}, Z_{1i}) = \psi_i^\top C^* C\, \psi_i - 2\langle C, \widehat{\xi}_i \otimes \psi_i\rangle_{\mathrm{HS}}.$$

By Lemma 3 at the scale $B_{\mathcal{F}}/\sqrt{n}$ for the empirical measure Q of $V_{11}, \dots, V_{1n}$, there is a finite $(B_{\mathcal{F}}/\sqrt{n}, 2)$-cover $\mathcal{N}_0$ of $\mathcal{F}^M$ in $L_2(Q)$ with $\log|\mathcal{N}_0| \le c_{\mathrm{FC}}\, WL \log W \log(4\sqrt{Mn})$. We denote by $\Pi_{B_\mathcal{F}} \colon \mathbb{R}^M \to \{u \in \mathbb{R}^M : \|u\| \le B_\mathcal{F}\}$ the projection onto the ball, and by $\mathcal{N} := \{\Pi_{B_\mathcal{F}} \circ \psi : \psi \in \mathcal{N}_0\}$ the set of the projections of the elements of $\mathcal{N}_0$. Since $\|\Pi_{B_\mathcal{F}}(u) - \Pi_{B_\mathcal{F}}(u')\| \le \|u - u'\|$ for $u, u' \in \mathbb{R}^M$ and $\Pi_{B_\mathcal{F}}(\psi(v)) = \psi(v)$ for every $\psi \in \mathcal{F}^M$, the set $\mathcal{N}$ is a $(B_\mathcal{F}/\sqrt{n}, 2)$-cover of $\mathcal{F}^M$ in $L_2(Q)$ with:

$$\log|\mathcal{N}| \le \log|\mathcal{N}_0| \le c_{\mathrm{FC}}\, WL \log W \log\Big(4\sqrt{Mn}\Big).$$

For any $\psi \in \mathcal{F}^M$, we choose $\psi' \in \mathcal{N}$ with $\frac{1}{n}\sum_{i=1}^n \|\psi_i - \psi_i'\|^2 \le B_\mathcal{F}^2/n$, where $\psi_i' := \psi'(V_{1i})$. For $\|C\|_{\mathrm{HS}} \le R_\lambda$, the matrix C^*C is symmetric with $\|C^*C\|_{\mathrm{op}} \le R_\lambda^2$. Thus, since $\|\psi_i\| \le B_\mathcal{F}$ and $\|\psi_i'\| \le B_\mathcal{F}$, we have:

$$\big|\psi_i^\top C^* C\, \psi_i - \psi_i'^\top C^* C\, \psi_i'\big| = \big|(\psi_i - \psi_i')^\top C^* C(\psi_i + \psi_i')\big| \le 2R_\lambda^2 B_\mathcal{F} \|\psi_i - \psi_i'\|.$$

By the Cauchy-Schwarz inequality, we also have $\big|\langle C(\psi_i - \psi_i'), \widehat{\xi}_i\rangle_{\mathcal{H}_\mathcal{Y}}\big| \le R_\lambda \sqrt{B_{\widehat{\xi}}} \|\psi_i - \psi_i'\|$. Since $\frac{1}{n}\sum_{i=1}^n \|\psi_i - \psi_i'\| \le B_\mathcal{F}/\sqrt{n}$ by the Cauchy-Schwarz inequality, every C with $\|C\|_{\mathrm{HS}} \le R_\lambda$ and every choice of ε_i's satisfy:

$$\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \big(\bar{\mathcal{L}}_{C\psi} - \bar{\mathcal{L}}_{C\psi'}\big)(V_{1i}, Z_{1i})\right| \le \Big(2R_\lambda^2 B_\mathcal{F} + 2R_\lambda \sqrt{B_{\widehat{\xi}}}\Big) \frac{B_\mathcal{F}}{\sqrt{n}}. \tag{63}$$

Next, we fix $\psi' \in \mathcal{N}$. Denote $S_1(\psi') := \frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi_i' \psi_i'^\top$ and $S_2(\psi') := \frac{1}{n}\sum_{i=1}^n \varepsilon_i\, \widehat{\xi}_i \otimes \psi_i'$. By the Cauchy-Schwarz inequality in the Frobenius and Hilbert-Schmidt inner products, and since $\|C^*C\|_F \le R_\lambda^2$ for $\|C\|_{\mathrm{HS}} \le R_\lambda$, we have:

$$\sup_{\|C\|_{\mathrm{HS}} \le R_\lambda} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\, \bar{\mathcal{L}}_{C\psi'}(V_{1i}, Z_{1i})\right| \le R_\lambda^2 \|S_1(\psi')\|_F + 2R_\lambda \|S_2(\psi')\|_{\mathrm{HS}}. \tag{64}$$

We now bound the two terms on the right-hand side. By Jensen's inequality and $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \ne j$, we have $\mathbb{E}_\varepsilon n\|S_1(\psi')\|_F \le (\sum_{i=1}^n \|\psi_i'\|^4)^{1/2} \le \sqrt{n} B_\mathcal{F}^2$. Moreover, denoting by $S_1^{(i)}(\psi')$

the matrix $S_1(\psi')$ with ε_i replaced by $-\varepsilon_i$, we have $\left|n\|S_1(\psi')\|_F - n\|S_1^{(i)}(\psi')\|_F\right| \le 2B_{\mathcal{F}}^2$. Thus, according to McDiarmid's inequality [BLM13, Theorem 6.2], every $t \in \mathbb{R}$ satisfies:

$$\mathbb{E}_\varepsilon[\exp\{t(n\|S_1(\psi')\|_F - \mathbb{E}_\varepsilon[n\|S_1(\psi')\|_F])\}] \le \exp\left(\frac{nB_{\mathcal{F}}^4 t^2}{2}\right).$$

Since $\|\widehat{\xi}_i \otimes \psi_i'\|_{\mathrm{HS}} \le \sqrt{B_{\widehat{\xi}}} B_{\mathcal{F}}$, we apply the same argument to $n\|S_2(\psi')\|_{\mathrm{HS}}$. The maximal inequality for $|\mathcal{N}|$ sub-Gaussian variables, which need not be independent [BLM13, Theorem 2.5], then yields:

$$\begin{aligned}
\mathbb{E}_\varepsilon \max_{\psi' \in \mathcal{N}} n\|S_1(\psi')\|_F &\le \sqrt{n} B_{\mathcal{F}}^2 \left(1 + \sqrt{2\log|\mathcal{N}|}\right), \\
\mathbb{E}_\varepsilon \max_{\psi' \in \mathcal{N}} n\|S_2(\psi')\|_{\mathrm{HS}} &\le \sqrt{n}\sqrt{B_{\widehat{\xi}}} B_{\mathcal{F}} \left(1 + \sqrt{2\log|\mathcal{N}|}\right).
\end{aligned} \tag{65}$$

Combining (63), (64) and (65), we obtain:

$$\mathrm{Rad}_n \le \left(R_\lambda^2 B_{\mathcal{F}}^2 + 2R_\lambda \sqrt{B_{\widehat{\xi}}} B_{\mathcal{F}}\right) \frac{1 + \sqrt{2\log|\mathcal{N}|}}{\sqrt{n}} + \left(2R_\lambda^2 B_{\mathcal{F}} + 2R_\lambda \sqrt{B_{\widehat{\xi}}}\right) \frac{B_{\mathcal{F}}}{\sqrt{n}}. \tag{66}$$

Finally, since $R_\lambda^2 = B_{\widehat{\xi}}/\lambda$ and $R_\lambda \sqrt{B_{\widehat{\xi}}} = B_{\widehat{\xi}}/\sqrt{\lambda} \le B_{\widehat{\xi}}/\lambda$ for $\lambda \le 1$, (66) and the bound on $\log|\mathcal{N}|$ yield:

$$\mathrm{Rad}_n \le \frac{B_{\widehat{\xi}}(B_{\mathcal{F}}^2 + 2B_{\mathcal{F}})}{\lambda} \cdot \frac{3 + \sqrt{2c_{\mathrm{FC}} W L \log W \log(4\sqrt{Mn})}}{\sqrt{n}}.$$

Since $n \ge 2$, $W \ge 2$ and c_{FC} is a universal constant, there is a universal constant $c > 0$ such that the numerator of the last fraction is at most $c\,(WL\log W \log(Mn))^{1/2}$. Inserting this bound into the right-hand side of (62) yields the claim with $c_{\mathrm{DF}} = 2cB_{\widehat{\xi}}(B_{\mathcal{F}}^2 + 2B_{\mathcal{F}})$. □

M.3 Proof of the Bound

We now begin the proof of the upper bound. We denote $\mu := \mu_{Y^1|V}$. The quantities λ and $\widetilde{\mu}_{Y^1|V}$ are as defined before Lemma 4. The norm $\|\cdot\|_{L_2(P_V)}$ is as defined in Appendix K.1.

Our main idea of the proof is to choose M auxiliary points to define a subspace on which to approximate the CCME. In particular, we construct M distinct points $\{\widetilde{y}_j\}_{j=1}^M \subset \mathcal{Y}$ whose fill distance $h = \sup_{y \in \mathcal{Y}} \min_j \|y - \widetilde{y}_j\|$ satisfies $h \asymp M^{-1/d_y}$ as follows: For $s \in (0, 1]$, let $\mathcal{T}_s \subseteq \mathcal{Y}$ be a maximal set with $\|y - y'\| > s$ for all distinct $y, y' \in \mathcal{T}_s$. By the maximality of $\mathcal{T}_s$, $\sup_{y \in \mathcal{Y}} \min_{y' \in \mathcal{T}_s} \|y - y'\| \le s$. Since the sets $\{y : \|y - y'\| < s/2\}$, $y' \in \mathcal{T}_s$, are disjoint and lie in a bounded neighborhood of $\mathcal{Y}$, we have $|\mathcal{T}_s| \le C_{\mathcal{Y}} s^{-d_y}$ for a constant $C_{\mathcal{Y}}$ that only depends on $\mathcal{Y}$. For any fixed choice of $M \ge C_{\mathcal{Y}}$, we take $s = (C_{\mathcal{Y}}/M)^{1/d_y}$ and add distinct points of $\mathcal{Y}$ to $\mathcal{T}_s$ until there are M points $\widetilde{y}_1, \ldots, \widetilde{y}_M$, which does not increase the fill distance. Thus, $h \le (C_{\mathcal{Y}}/M)^{1/d_y}$. Conversely, since $\mathcal{Y} \subseteq \bigcup_{j=1}^M \{y : \|y - \widetilde{y}_j\| \le h\}$, we have $h \gtrsim M^{-1/d_y}$.

By Assumption 7 and Proposition 2, the features $\phi(\widetilde{y}_1), \ldots, \phi(\widetilde{y}_M)$ are linearly independent and their Gram matrix $\mathbf{K}_M = (k_{\mathcal{Y}}(\widetilde{y}_j, \widetilde{y}_l))_{j,l=1}^M$ is invertible. Consequently, the subspace $\mathcal{H}_M := \mathrm{span}\{\phi(\widetilde{y}_j)\}_{j=1}^M$ satisfies $\dim \mathcal{H}_M = M$. We denote by $\Pi_M \colon \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ the orthogonal projection onto $\mathcal{H}_M$. Let $\{e_k\}_{k=1}^M$ be an orthonormal basis of $\mathcal{H}_M$. We use the isometry $C_M : \mathbb{R}^M \to \mathcal{H}_M$ defined by $C_M(w) := \sum_{k=1}^M w_k e_k$, for which $\|C_M\|_{\mathrm{HS}}^2 = M$. Let $\theta_c \in \Theta$ be a parameter, and denote $\psi_c := \psi_{\theta_c}$. We define the approximation, nuisance and projection errors:

$$\mathbf{B} := \mathbb{E}\big[\|C_M \psi_c(V) - \Pi_M \mu(V)\|_{\mathcal{H}_{\mathcal{Y}}}^2\big], \qquad \mathbf{C} := \mathbb{E}\big[\|\widetilde{\mu}_{Y^1|V}(V) - \mu(V)\|_{\mathcal{H}_{\mathcal{Y}}}^2\big],$$
$$\mathbf{D} := \mathbb{E}\big[\|\Pi_M \mu(V) - \mu(V)\|_{\mathcal{H}_{\mathcal{Y}}}^2\big].$$

Then, we have:

$$\begin{aligned}
\|C_M \psi_c - \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 &\lesssim \mathbf{B} + \mathbf{D} + \|\widetilde{\mu}_{Y^1|V} - \mu\|_{L_2(P_V)}^2, \\
\|\widehat{\mu}_{\mathrm{DF}} - \mu\|_{L_2(P_V)}^2 &\lesssim \|\widehat{\mu}_{\mathrm{DF}} - \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 + \|\widetilde{\mu}_{Y^1|V} - \mu\|_{L_2(P_V)}^2.
\end{aligned}$$

Since θ_c and C_M do not depend on $\mathcal{D}_1$, Lemma 4 with $C_c = C_M$, the identity $\|C_M\|_{\mathrm{HS}}^2 = M$ and the last display yield, after we take the expectation over $\mathcal{D}_0$, the following bound for $\lambda \le 1$ and $n \ge 2$:

$$\mathbb{E}\big[\|\widehat{\mu}_{\mathrm{DF}}(V) - \mu(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D} + \lambda M, \tag{67}$$

where $\mathbf{A} := \lambda^{-1}(WL \log W \log(Mn)/n)^{1/2}$ is the statistical error. It remains to bound $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$.

Term B: The Approximation Error. Recall that $C_M\psi(v) = \sum_{k=1}^M \psi_k(v)e_k$. Thus, for $C_M\psi(v)$ to approximate $\Pi_M\mu_{Y^1|V}(v) = \sum_{k=1}^M c_k(v)e_k$, with c_k as in Lemma 10, the neural network $\psi \in \mathcal{F}^M$ must approximate the coefficient vector $\mathbf{c}(v) := (c_1(v), \dots, c_M(v))$. To achieve this, we use Lemma 11, which provides an approximation of $\Pi_M\mu_{Y^1|V}$ by a fully connected network with M bounded outputs. The conditions of Lemma 11 are that (i) Assumptions 2, 4 and 6 hold, and (ii) the smallest hidden width $D := \min_{1\le i\le L-1} d_i$ satisfies $D \ge \bar{w}M$ for a constant $\bar{w} \ge 2$ defined in Lemma 11.

We now check these conditions. Condition (i) is immediate by our assumptions. Thus, it remains to show that condition (ii) holds for all sufficiently large n. Our choices of M and WL in Case 1 of the theorem yield $WL/M^2 \asymp n^{a/(4ab+4b+a)}$ and those in Case 2 yield $WL/M^2 \asymp n^{1/(4b+1)}(\log n)^{-4d_y}$. Since L does not depend on n, we have $W/M^2 \to \infty$ as $n \to \infty$ in both cases. By the assumption $\min_{1\le i\le L-1} d_i \asymp \max_{1\le i\le L-1} d_i$, every hidden width d_i satisfies $d_i \lesssim D$. Thus, $W = \sum_{i=1}^L (d_i d_{i-1} + d_i) \lesssim LD^2 + MD$, which implies $W/M^2 \lesssim L(D/M)^2 + D/M$. Since $W/M^2 \to \infty$, we must have $D/M \to \infty$. We therefore conclude that condition (ii) holds for all sufficiently large n. Consequently, Lemma 11 yields constants $\bar{B} > 0$ and L_0 such that, for any $B_\mathcal{F}^2 \ge \bar{B}$, any $L \ge \max\{L_0, 3\}$ and all sufficiently large n, there exists $\psi_c \in \mathcal{F}^M = \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})$ such that $\mathbb{E}[\|C_M\psi_c(V) - \Pi_M\mu(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim (M/(DL))^{2b}$. Since $M \lesssim D$, the bound $W \lesssim LD^2 + MD$ gives $W \lesssim LD^2$. Thus, inserting ψ_c into the definition of $\mathbf{B}$ yields:

$$\mathbf{B} = \mathbb{E}\big[\|C_M\psi_c(V) - \Pi_M\mu(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim \left(\frac{M}{DL}\right)^{2b} \lesssim \left(\frac{M^2}{WL}\right)^{b}.$$

Term C: The Nuisance Error. Since the assumptions of the theorem include $\pi(x), \widehat{\pi}(x) \in [\varepsilon, 1)$ and $\sup_{x\in\mathcal{X}}\|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \le B_{\mu_0}$, we can use Lemma 1 to obtain:

$$\mathbf{C} \lesssim \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.$$

Term D: The Projection Error. By Assumptions 4 and 6, we have $\mathbb{E}[\|\mathfrak{p}^1(\cdot|V)\|_{L^2(\mathcal{Y})}^2] \le \sup_{v\in\mathcal{V}} p_V(v) \int_\mathcal{Y} \|\mathfrak{p}^1(y|\cdot)\|_{L^2(\mathcal{V})}^2 \,\mathrm{d}y < \infty$, so $\mathfrak{p}^1(\cdot|V) \in L^2(\mathcal{Y})$ almost surely. According to Lemma 12, the conditions that (i) the auxiliary points have fill distance $h \asymp M^{-1/d_y}$, (ii) Π_M is the orthogonal projection onto $\mathrm{span}\{\phi(\widetilde{y}_j)\}_{j=1}^M$ and (iii) $\mathfrak{p}^1(\cdot|V) \in L^2(\mathcal{Y})$ almost surely yield, under Assumptions 4 and 5:

$$\mathbf{D} = \mathbb{E}\big[\|\Pi_M\mu_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim M^{-2\tau/d_y}\,\mathbb{E}\Big[\|\mathfrak{p}^1(\cdot|V)\|_{L^2(\mathcal{Y})}^2\Big] \lesssim M^{-a}.$$

In Case 2 of the theorem, the second part of Lemma 12 yields, for some $B_0 > 0$:

$$\mathbf{D} \lesssim \exp\Big(-B_0 M^{\frac{1}{2d_y}} \log M\Big). \tag{68}$$

Combining the bounds. In Case 1 of the theorem, we obtain the following inequality from (67) and the bounds of Terms $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$:

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{DF}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim \left(\frac{M^2}{WL}\right)^{b} + M^{-a} + \lambda M + \frac{1}{\lambda}\left(\frac{WL\log W\log(Mn)}{n}\right)^{1/2} + \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.$$

Recall our choices of $M \asymp n^{\frac{b}{4ab+4b+a}}$ and $WL \asymp n^{\frac{a+2b}{4ab+4b+a}}$. By choosing $\lambda \asymp n^{-\frac{b(a+1)}{4ab+4b+a}}$, we have $(M^2/(WL))^b \asymp M^{-a} \asymp \lambda M \asymp \lambda^{-1}(WL/n)^{1/2} \asymp n^{-\frac{ab}{4ab+4b+a}}$ and the conditions of (67) hold for all sufficiently large n. Therefore, we obtain the bound:

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{DF}}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}] \lesssim_{\log n} n^{-\frac{ab}{4ab+4b+a}} + \min\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\}.$$

In Case 2, we instead obtain from (67), the bounds of Terms **B** and **C**, and (68) that:

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\mu}_{\mathrm{DF}}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}] \\
&\lesssim \left(\frac{M^2}{WL}\right)^b + \exp\left(-B_0 M^{\frac{1}{2d_y}} \log M\right) + \lambda M \\
&\quad + \frac{1}{\lambda}\left(\frac{WL \log W \log(Mn)}{n}\right)^{1/2} + \min\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\}.
\end{aligned}$$

Recall our choices of $M \asymp (\log n)^{2d_y}$ and $WL \asymp n^{\frac{1}{4b+1}}$. Since $M^{\frac{1}{2d_y}} \gtrsim \log n$ and $\log M \to \infty$, we have $\exp(-B_0 M^{\frac{1}{2d_y}} \log M) \le n^{-1}$ for all sufficiently large n. By choosing $\lambda \asymp n^{-\frac{b}{4b+1}}$, we have $(M^2/(WL))^b + \lambda M + (1/\lambda)(WL/n)^{1/2} \lesssim_{\log n} n^{-\frac{b}{4b+1}}$. Moreover, the conditions of (67) hold for all sufficiently large n. Therefore, we obtain the bound:

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{DF}}(V) - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}] \lesssim_{\log n} n^{-\frac{b}{4b+1}} + \min\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\}.$$

■

N Proof of the Upper Bound for the Neural-Kernel Estimator (Theorem 4)

N.1 Notations and Setup

We recall the definition of the Neural-Kernel estimator. We consider the M grid points $\{\widetilde{y}_j\}_{j=1}^M \subset \mathcal{Y}$ of Assumption 8.

Let $f_\theta : \mathcal{V} \to \mathbb{R}^M$ be a neural network parameterized by $\theta \in \Theta$ of Theorem 4 with M outputs. We denote by $\mathcal{F}_{\mathrm{FC}}$ the class of neural networks f_θ with the widths $\mathbf{d} = (d_v, d_1, \ldots, d_{L-1}, d_L)$ where $d_L = M$, that is, $\mathcal{F}_{\mathrm{FC}} := \{f_\theta : \theta \in \prod_{i=1}^L (\mathbb{R}^{d_i \times d_{i-1}} \times \mathbb{R}^{d_i}),\ \sup_{v \in \mathcal{V}} \|\sum_{j=1}^M f_\theta(v)_j \phi(\widetilde{y}_j)\|_{\mathcal{H}_\mathcal{Y}} \le B_\mathcal{F}\}$. We denote by $\mathcal{F}_{\mathrm{NK}}$ the class of the estimator: $\mathcal{F}_{\mathrm{NK}} := \{v \mapsto \sum_{j=1}^M f_j(v)\phi(\widetilde{y}_j) : f \in \mathcal{F}_{\mathrm{FC}}\}$. Thus, the elements of $\mathcal{F}_{\mathrm{NK}}$ take values in $\mathcal{H}_M := \mathrm{span}\{\phi(\widetilde{y}_j)\}_{j=1}^M$.

N.2 Preliminary Lemmas

The following lemma provides an orthonormal basis of $\mathcal{H}_M$ in which the Neural-Kernel estimator has network coordinates:

Lemma 5. *Let Assumption 7 hold, and denote $e_k := \sum_{j=1}^M (\mathbf{K}_M^{-1/2})_{jk} \phi(\widetilde{y}_j)$ for $k = 1, \ldots, M$. Then, $\{e_k\}_{k=1}^M$ is an orthonormal basis of $\mathcal{H}_M$. In addition, every $w \in \mathbb{R}^M$ satisfies*

$$\sum_{j=1}^M w_j \phi(\widetilde{y}_j) = \sum_{k=1}^M (\mathbf{K}_M^{1/2} w)_k \, e_k.$$

Proof. By Assumption 7, φ is strictly positive definite [Wen04, Chapter 6], and thus $\mathbf{K}_M$ is invertible by Proposition 2. Since the Gram matrix of $\{e_k\}_{k=1}^M$ is $\mathbf{K}_M^{-1/2}\mathbf{K}_M\mathbf{K}_M^{-1/2} = \mathbf{I}_M$, the set $\{e_k\}_{k=1}^M$ is an orthonormal basis of $\mathcal{H}_M$. For the second claim, we expand the right-hand side:

$$\sum_{k=1}^M (\mathbf{K}_M^{1/2} w)_k \, e_k = \sum_{j=1}^M (\mathbf{K}_M^{-1/2}\mathbf{K}_M^{1/2} w)_j \, \phi(\widetilde{y}_j) = \sum_{j=1}^M w_j \phi(\widetilde{y}_j).$$

□

In the Deep Feature estimator, the map C is an element of a ball in the infinite-dimensional space $\mathcal{BL}(\mathbb{R}^M; \mathcal{H}_\mathcal{Y})$, for which we have no entropy bound. Thus, we bound the supremum over C in closed form in (64), with which we can employ the Rademacher complexity argument. The class of the Neural-Kernel estimator, on the other hand, consists of networks with finitely many parameters, for which Lemma 3 provides an entropy bound. Thus, we can employ the following oracle inequality to achieve, up to logarithmic factors, an excess risk bound of order $O(n^{-1})$ instead of the order $O(n^{-1/2})$ that we obtained for the Deep Feature estimator. Since [vdVW23, Theorem 3.5.4] is stated for multiple models in the context of model selection, we adapt this theorem to a single model:

Lemma 6 (Oracle Inequality [vdVW23, Theorem 3.5.4]). *Let $\mathcal{G} \subseteq (\mathcal{V} \times \mathcal{Z} \to \mathbb{R})$ be a class of functions that do not depend on $\mathcal{D}_1$. Let $g_0 \colon \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$ be a function that does not depend on $\mathcal{D}_1$, and let $\widehat{g} \in \arg\min_{g \in \mathcal{G}} \mathbb{E}_n[g(V,Z)]$. Assume that (i) every $g \in \mathcal{G}$ satisfies $|g| \leq 1$, (ii) every $g \in \mathcal{G}$ satisfies $\mathbb{E}[(g(V,Z) - g_0(V,Z))^2 \mid \mathcal{D}_0] \leq \mathbb{E}[g(V,Z) - g_0(V,Z) \mid \mathcal{D}_0]$, and (iii) there exist $\delta_n > 0$ and a function $\vartheta \colon (0,\infty) \to [0,\infty)$ such that $\delta \mapsto \vartheta(\delta)/\delta$ is nonincreasing, $\vartheta(\delta_n) \leq \sqrt{n}\,\delta_n^2$ and, for every $\delta \geq \delta_n$,*

$$\mathbb{E}\Big[\sup \Big\{ \sqrt{n} \left(\mathbb{E}_n[g(V,Z) - g'(V,Z)] - \mathbb{E}[g(V,Z) - g'(V,Z) \mid \mathcal{D}_0] \right) : g, g' \in \mathcal{G},\ \mathbb{E}[(g(V,Z) - g'(V,Z))^2 \mid \mathcal{D}_0] < \delta^2 \Big\} \,\Big|\, \mathcal{D}_0 \Big] \leq \vartheta(\delta).$$

Then, there exists a universal constant $c_5 > 0$ such that

$$\mathbb{E}[\widehat{g}(V,Z) - g_0(V,Z) \mid \mathcal{D}_0] \leq c_5 \inf_{g \in \mathcal{G}} \mathbb{E}[g(V,Z) - g_0(V,Z) \mid \mathcal{D}_0] + c_5 \left(\delta_n^2 + \frac{1}{n} \right).$$

In condition (iii) of Lemma 6, we need a bound on the expected supremum of the empirical process over the pairs $g, g' \in \mathcal{G}$ with $\mathbb{E}[(g(V,Z) - g'(V,Z))^2 \mid \mathcal{D}_0] < \delta^2$. For a class of bounded functions with small second moments, the following local maximal inequality provides such a bound in terms of the uniform entropy integral of the class:

Lemma 7 (Local Maximal Inequality [vdVW23, Theorem 2.14.2]). *With the same notations as Lemma 6, let $\delta \in (0,1)$, and assume that (i) every $g \in \mathcal{G}$ satisfies $|g| \leq 1$ and (ii) every $g \in \mathcal{G}$ satisfies $\mathbb{E}[g(V,Z)^2 \mid \mathcal{D}_0] < \delta^2$. Denote by $\mathcal{J}_\mathcal{G}(\delta)$ the following uniform entropy integral of $\mathcal{G}$, in which the supremum is taken over all probability measures Q on $\mathcal{V} \times \mathcal{Z}$:*

$$\mathcal{J}_\mathcal{G}(\delta) \coloneqq \int_0^\delta \sup_Q \sqrt{1 + \mathcal{H}_2(\mathcal{G}, \varepsilon, Q)}\, \mathrm{d}\varepsilon. \tag{69}$$

Then, there exists a universal constant $c_6 > 0$ such that

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} \sqrt{n}\, |\mathbb{E}_n[g(V,Z)] - \mathbb{E}[g(V,Z) \mid \mathcal{D}_0]| \,\middle|\, \mathcal{D}_0 \right] \leq c_6 \mathcal{J}_\mathcal{G}(\delta) \left(1 + \frac{\mathcal{J}_\mathcal{G}(\delta)}{\delta^2 \sqrt{n}} \right).$$

As a result, we obtain the following excess risk bound for vector-valued functions:

Lemma 8 (Excess Risk Bound). *Let $\Pi_M \colon \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ be the orthogonal projection onto an M-dimensional subspace $\mathcal{H}_M \subseteq \mathcal{H}_\mathcal{Y}$. Let $\{e_k\}_{k=1}^M$ be an orthonormal basis for $\mathcal{H}_M$. Let $\mathcal{F}^M \subseteq (\mathcal{V} \to \mathbb{R}^M)$ be a class of functions, identified with a subset of $(\mathcal{V} \to \mathcal{H}_M)$ through the isometry $\iota(w) \coloneqq \sum_{k=1}^M w_k e_k$ such that $\sup_{f \in \mathcal{F}^M} \sup_{v \in \mathcal{V}} \|f(v)\|_{\mathcal{H}_\mathcal{Y}} \leq B$ for some $B > 0$. Let $\widehat{f}$ and $f^\star$ be the minimizers in (54) with this isometry, and denote $\widetilde{\mu}_{Y^1|V}(v) \coloneqq \mathbb{E}[\widehat{\xi}(Z) | V = v, \mathcal{D}_0]$. In addition, assume that (i) there exist $D_\mathcal{F}, A_\mathcal{F} \geq 1$ such that, for every probability measure $Q_\mathcal{V}$ on $\mathcal{V}$ and every $\varepsilon \in (0, B]$,*

$$\mathcal{H}_2(\mathcal{F}^M, \varepsilon, Q_\mathcal{V}) \leq D_\mathcal{F} \log(A_\mathcal{F} B/\varepsilon), \tag{70}$$

and (ii) $\|\widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \leq B'$ for some constant $B' > 0$ and all z. Then, there exists a universal constant $c_3 > 0$ such that, for any $n \geq 2$,

$$\mathbb{E}\left[\|\widehat{f} - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 \,\middle|\, \mathcal{D}_0 \right] \leq c_3 \|f^\star - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 + \frac{c_3 (B + B')^2 D_\mathcal{F} \log(A_\mathcal{F} n)}{n}. \tag{71}$$

Proof of Lemma 8. We denote $f_0 \coloneqq \Pi_M \widetilde{\mu}_{Y^1|V}$ and $\mathfrak{L} \coloneqq B + 3B'$. Our main idea of the proof is to apply Lemma 6 to the class $\mathcal{G} \coloneqq \{\mathcal{L}_f/\mathfrak{L}^2 : f \in \mathcal{F}^M\}$ and the function $g_0 \coloneqq \mathcal{L}_{f_0}/\mathfrak{L}^2$. We first show that conditions (i) and (ii) of Lemma 6 hold. For every $f \in \mathcal{F}^M$, both $f(v)$ and $f_0(v)$ belong to $\mathcal{H}_M$, so we have:

$$\mathcal{L}_f(v,z) - \mathcal{L}_{f_0}(v,z) = \langle f(v) - f_0(v), f(v) + f_0(v) - 2\Pi_M \widehat{\xi}(z)\rangle_{\mathcal{H}_\mathcal{Y}}. \tag{72}$$

Since the conditional expectation commutes with the bounded linear operator Π_M, the definitions of $\widetilde{\mu}_{Y^1|V}$ and f_0 yield $\mathbb{E}[\Pi_M \widehat{\xi}(Z)|V, \mathcal{D}_0] = \Pi_M \widetilde{\mu}_{Y^1|V}(V) = f_0(V)$. Taking the expectation of the last display conditionally on V and $\mathcal{D}_0$ therefore yields:

$$\mathbb{E}[\mathcal{L}_f(V,Z) - \mathcal{L}_{f_0}(V,Z) \mid \mathcal{D}_0] = \|f - f_0\|^2_{L_2(P_V)}. \tag{73}$$

Next, we derive a uniform bound on $\mathcal{L}_f$ and a bound on the second moment of $\mathcal{L}_f - \mathcal{L}_{f_0}$. For all v and z, we have $\|f(v)\|_{\mathcal{H}_\mathcal{Y}} \le B$ and $\|\widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \le B'$, which imply $0 \le \mathcal{L}_f(v,z) \le (B + B')^2 \le \mathfrak{L}^2$. Thus, condition (i) of Lemma 6 holds. Moreover, the bound on $\widehat{\xi}$ implies $\sup_{v\in\mathcal{V}} \|\widetilde{\mu}_{Y^1|V}(v)\|_{\mathcal{H}_\mathcal{Y}} \le B'$. Since $\|\Pi_M\|_{\mathrm{op}} \le 1$, the bounds on $\widetilde{\mu}_{Y^1|V}$ and $\widehat{\xi}$ imply $\|f_0(v)\|_{\mathcal{H}_\mathcal{Y}} \le B'$ and $\|\Pi_M \widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \le B'$. By (72) and the Cauchy-Schwarz inequality, the bounds on f, f_0 and $\Pi_M \widehat{\xi}$ imply $|\mathcal{L}_f(v,z) - \mathcal{L}_{f_0}(v,z)| \le \mathfrak{L}\|f(v) - f_0(v)\|_{\mathcal{H}_\mathcal{Y}}$. Squaring this bound and using (73), we obtain:

$$\mathbb{E}[(\mathcal{L}_f(V,Z) - \mathcal{L}_{f_0}(V,Z))^2 \mid \mathcal{D}_0] \le \mathfrak{L}^2\, \mathbb{E}[\mathcal{L}_f(V,Z) - \mathcal{L}_{f_0}(V,Z) \mid \mathcal{D}_0].$$

Thus, condition (ii) of Lemma 6 holds. It remains to check condition (iii) of Lemma 6, which will be done via a covering argument combined with Lemma 7.

Let $t \in (0, B]$, and let Q be a probability measure on $\mathcal{V} \times \mathcal{Z}$. By (70) with $Q_\mathcal{V}$ as the marginal of Q on $\mathcal{V}$, there is a $(t,2)$-cover $\mathcal{N}_0$ of $\mathcal{F}^M$ in $L_2(Q_\mathcal{V})$ with $\log|\mathcal{N}_0| \le D_\mathcal{F} \log(A_\mathcal{F} B/t)$. Consider the ball $\mathcal{B}_M(B) \coloneqq \{u \in \mathcal{H}_M : \|u\|_{\mathcal{H}_\mathcal{Y}} \le B\} \subset \mathcal{H}_M$ and denote by $\Pi_B \colon \mathcal{H}_M \to \mathcal{B}_M(B)$ the projection onto the ball, and let $\mathcal{N} \coloneqq \{\Pi_B \circ h : h \in \mathcal{N}_0\}$. Since $\|\Pi_B(u) - \Pi_B(u')\|_{\mathcal{H}_\mathcal{Y}} \le \|u - u'\|_{\mathcal{H}_\mathcal{Y}}$ for $u, u' \in \mathcal{H}_M$ and $\Pi_B(f(v)) = f(v)$ for every $f \in \mathcal{F}^M$, the set $\mathcal{N}$ is also a $(t,2)$-cover of $\mathcal{F}^M$ in $L_2(Q_\mathcal{V})$.

For maps $f, f' \colon \mathcal{V} \to \mathcal{H}_M$ with values in $\mathcal{B}_M(B)$, the identity (72) with f_0 replaced by f' and the Cauchy-Schwarz inequality yield:

$$|\mathcal{L}_f(v,z) - \mathcal{L}_{f'}(v,z)| \le 2\mathfrak{L}\|f(v) - f'(v)\|_{\mathcal{H}_\mathcal{Y}}.$$

Therefore, the set $\mathcal{N}_\mathcal{L} \coloneqq \{\mathcal{L}_h/\mathfrak{L}^2 : h \in \mathcal{N}\} \subseteq (\mathcal{V} \times \mathcal{Z} \to \mathbb{R})$ forms a $(2t/\mathfrak{L}, 2)$-cover of $\mathcal{G}$ in $L_2(Q)$ with at most $|\mathcal{N}_0|$ elements. In particular, given $\varepsilon \in (0,1]$, we choose $t = \min\{\mathfrak{L}\varepsilon, B\}$ so that the set $\mathcal{N}_\mathcal{L}$ is a $(2\varepsilon, 2)$-cover of $\mathcal{G}$ in $L_2(Q)$ with $\log|\mathcal{N}_0| \le D_\mathcal{F} \log(A_\mathcal{F}/\varepsilon)$.

Let $\delta' \in (0,4)$ and denote by $\mathcal{K} \subseteq (\mathcal{V} \times \mathcal{Z} \to \mathbb{R})$ the class of the functions $(g - g')/4$ with $g, g' \in \mathcal{G}$ and $\mathbb{E}[(g(V,Z) - g'(V,Z))^2 \mid \mathcal{D}_0] < \delta'^2$. It is easy to verify that the conditions (i) and (ii) of Lemma 7 with $\delta = \delta'/4$ are satisfied for the class $\mathcal{K}$. In addition, the set $\{(h - h')/4 : h, h' \in \mathcal{N}_\mathcal{L}\}$ forms an $(\varepsilon, 2)$-cover of $\mathcal{K}$ in $L_2(Q)$ with at most $|\mathcal{N}_0|^2$ elements. Consequently, $\mathcal{H}_2(\mathcal{K}, \varepsilon, Q) \le 2\log|\mathcal{N}_0| \le 2D_\mathcal{F} \log(A_\mathcal{F}/\varepsilon)$. Therefore, the uniform entropy integral $\mathcal{J}_\mathcal{K}(\delta'/4)$ defined in (69) satisfies:

$$\mathcal{J}_\mathcal{K}(\delta'/4) \le \int_0^{\delta'/4} \sqrt{1 + 2D_\mathcal{F} \log \frac{A_\mathcal{F}}{\varepsilon}}\, \mathrm{d}\varepsilon \le \frac{\delta'}{\sqrt{2}} \sqrt{D_\mathcal{F} \log \frac{4eA_\mathcal{F}}{\delta'}}.$$

For $\delta' \ge 1/\sqrt{n}$, the right-hand side of the last display is at most $(\delta'/\sqrt{2})\sqrt{D_\mathcal{F} \log(4eA_\mathcal{F} n)}$. Thus, according to Lemma 7 for the class $\mathcal{K}$ with $\delta = \delta'/4$, this bound and conditions (i) and (ii) of this lemma yield a universal constant $c_4 \ge 1$ such that, for $1/\sqrt{n} \le \delta' < 4$,

$$\begin{aligned}
&\mathbb{E}\Big[\sup \big\{ \sqrt{n}\, \big(\mathbb{E}_n[g(V,Z) - g'(V,Z)] - \mathbb{E}[g(V,Z) - g'(V,Z) \mid \mathcal{D}_0]\big) : \\
&\qquad g, g' \in \mathcal{G},\ \mathbb{E}[(g(V,Z) - g'(V,Z))^2 \mid \mathcal{D}_0] < \delta'^2 \big\} \,\Big|\, \mathcal{D}_0 \Big] \\
&\le c_4 \left(\delta' \sqrt{D_\mathcal{F} \log(4eA_\mathcal{F} n)} + \frac{D_\mathcal{F} \log(4eA_\mathcal{F} n)}{\sqrt{n}} \right).
\end{aligned}$$

Since $0 \le g \le 1$ for every $g \in \mathcal{G}$, the supremum on the left-hand side of the last display for $\delta' > 2$ is the same as the supremum for $\delta' = 2$. Moreover, the right-hand side of the last display is nondecreasing in δ'. Thus, condition (iii) in Lemma 6 holds with $\vartheta(\delta') = c_4(\delta'\sqrt{D_{\mathcal{F}} \log(4eA_{\mathcal{F}}n)} + D_{\mathcal{F}} \log(4eA_{\mathcal{F}}n)/\sqrt{n})$ and $\delta_n = (2c_4 + 1)\sqrt{D_{\mathcal{F}} \log(4eA_{\mathcal{F}}n)/n}$, which satisfies $\delta_n \ge 1/\sqrt{n}$ and $\vartheta(\delta_n) \le \sqrt{n}\,\delta_n^2$.

By (73), the law of iterated expectations, conditionally on $\mathcal{D}_1$, and Lemma 6, we obtain:

$$\begin{aligned}
\mathbb{E}[\|\widehat{f} - f_0\|_{L_2(P_V)}^2 \mid \mathcal{D}_0] &= \mathbb{E}[\mathcal{L}_{\widehat{f}}(V, Z) - \mathcal{L}_{f_0}(V, Z) \mid \mathcal{D}_0] \\
&\le c_5 \inf_{f \in \mathcal{F}^M} \mathbb{E}[\mathcal{L}_f(V, Z) - \mathcal{L}_{f_0}(V, Z) \mid \mathcal{D}_0] + c_5 \mathfrak{L}^2 \left(\delta_n^2 + \frac{1}{n}\right) \\
&= c_5 \|f^\star - f_0\|_{L_2(P_V)}^2 + c_5 \mathfrak{L}^2 \left(\delta_n^2 + \frac{1}{n}\right).
\end{aligned}$$

Since $\mathfrak{L} \le 3(B + B')$, $D_{\mathcal{F}}, A_{\mathcal{F}} \ge 1$ and $n \ge 2$, we obtain (71) for a universal constant c_3. □

N.3 Proof of the Bound

We now set up for the proof of the upper bound. Let $\mu_{\widehat{\xi}}^\star := \arg\min_{\mu \in \mathcal{F}_{\mathrm{NK}}} \mathbb{E}[\|\mu(V) - \widehat{\xi}(Z)\|_{\mathcal{H}_\mathcal{Y}}^2 | \mathcal{D}_0]$ be the best-in-class estimator. Let $\Pi_M : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ be the orthogonal projection operator onto the subspace $\mathcal{H}_M$ and $\widetilde{\mu}_{Y^1|V}(v) := \mathbb{E}[\widehat{\xi}(Z)|V = v, \mathcal{D}_0]$. By the triangle inequality:

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mu}_{\mathrm{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim\ & \mathbb{E}\big[\|\widehat{\mu}_{\mathrm{NK}}(V) - \Pi_M \widetilde{\mu}_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \\
&+ \mathbb{E}\big[\|\Pi_M \widetilde{\mu}_{Y^1|V}(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \\
&+ \mathbb{E}\big[\|\Pi_M \mu_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big].
\end{aligned} \tag{74}$$

Let $c_3 > 0$ be the universal constant in the bound (71) of Lemma 8. We define the statistical error:

$$\mathbf{A} := \mathbb{E}\Big[\|\widehat{\mu}_{\mathrm{NK}} - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 - c_3 \|\mu_{\widehat{\xi}}^\star - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2\Big].$$

Let $\mu \in \mathcal{F}_{\mathrm{NK}}$. Since the conditional expectation commutes with Π_M, we have $\mathbb{E}[\Pi_M \widehat{\xi}(Z)|V, \mathcal{D}_0] = \Pi_M \widetilde{\mu}_{Y^1|V}(V)$. This identity and the fact that $\mu - \Pi_M \widetilde{\mu}_{Y^1|V}$ takes values in $\mathcal{H}_M$ imply that the cross term in the following expansion vanishes:

$$\begin{aligned}
\mathbb{E}[\|\mu(V) - \widehat{\xi}(Z)\|_{\mathcal{H}_\mathcal{Y}}^2 | \mathcal{D}_0] &= \mathbb{E}[\|\mu(V) - \Pi_M \widetilde{\mu}_{Y^1|V}(V) + \Pi_M \widetilde{\mu}_{Y^1|V}(V) - \widehat{\xi}(Z)\|_{\mathcal{H}_\mathcal{Y}}^2 | \mathcal{D}_0] \\
&= \|\mu - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 + \mathbb{E}[\|\Pi_M \widetilde{\mu}_{Y^1|V}(V) - \widehat{\xi}(Z)\|_{\mathcal{H}_\mathcal{Y}}^2 | \mathcal{D}_0].
\end{aligned}$$

Since the last term on the right-hand side does not depend on μ, the definition of $\mu_{\widehat{\xi}}^\star$ yields $\mu_{\widehat{\xi}}^\star \in \arg\min_{\mu \in \mathcal{F}_{\mathrm{NK}}} \|\mu - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2$. Consequently,

$$\begin{aligned}
\|\mu_{\widehat{\xi}}^\star - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 &= \inf_{\mu \in \mathcal{F}_{\mathrm{NK}}} \|\mu - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 \\
&\lesssim \inf_{\mu \in \mathcal{F}_{\mathrm{NK}}} \|\mu - \Pi_M \mu_{Y^1|V}\|_{L_2(P_V)}^2 + \|\Pi_M \mu_{Y^1|V} - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2.
\end{aligned}$$

We insert the definition of $\mathbf{A}$ and the last display into the right-hand side of (74). Taking the expectation over $\mathcal{D}_0$ then yields the following decomposition, with a constant $c > 0$ that only depends on c_3:

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\mu}_{\mathrm{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \\
&\quad \le 3 \underbrace{\mathbb{E}\Big[\|\widehat{\mu}_{\mathrm{NK}} - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2 - c_3 \|\mu_{\widehat{\xi}}^\star - \Pi_M \widetilde{\mu}_{Y^1|V}\|_{L_2(P_V)}^2\Big]}_{\mathbf{A}\text{: Statistical Error}} \\
&\qquad + c \underbrace{\inf_{\mu \in \mathcal{F}_{\mathrm{NK}}} \mathbb{E}\big[\|\mu(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big]}_{\mathbf{B}\text{: Approximation Error}} \\
&\qquad + c \underbrace{\mathbb{E}\big[\|\Pi_M \widetilde{\mu}_{Y^1|V}(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big]}_{\mathbf{C}\text{: Nuisance Error}} \\
&\qquad + c \underbrace{\mathbb{E}\big[\|\Pi_M \mu_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big]}_{\mathbf{D}\text{: Projection Error}}.
\end{aligned} \tag{75}$$

We now bound each of these terms.

Term A: The Statistical Error. The excess risk bound of Lemma 8 requires an isometry from $\mathbb{R}^M$ onto $\mathcal{H}_M$. Since the map $w \mapsto \sum_{j=1}^M w_j \phi(\widetilde{y}_j)$ is not an isometry, we first change the coordinates of the networks. We write

$$\widehat{\mu}_{\mathrm{NK}}(v) = \sum_{j=1}^M \widehat{f}_j(v)\phi(\widetilde{y}_j), \qquad \mu^\star_{\widehat{\xi}}(v) = \sum_{j=1}^M f^\star_j(v)\phi(\widetilde{y}_j),$$

where $f^\star \in \mathcal{F}_{\mathrm{FC}}$ and $\widehat{f} := f_{\widehat{\theta}}$ for the minimizer $\widehat{\theta}$ of (9) over Θ. By Lemma 5, every $f \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, \infty)$ satisfies $\sum_{j=1}^M f_j(v)\phi(\widetilde{y}_j) = \sum_{k=1}^M \psi_k(v) e_k$ with $\psi := \mathbf{K}_M^{1/2} f$.

For $f \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, \infty)$ with the output layer $(\mathbf{W}_L, \mathfrak{b}_L)$, the map $\psi = \mathbf{K}_M^{1/2} f$ is the network with the same hidden layers as f and the output layer $(\mathbf{K}_M^{1/2}\mathbf{W}_L, \mathbf{K}_M^{1/2}\mathfrak{b}_L)$. Since $\mathbf{K}_M$ is invertible, the map $(\mathbf{W}_L, \mathfrak{b}_L) \mapsto (\mathbf{K}_M^{1/2}\mathbf{W}_L, \mathbf{K}_M^{1/2}\mathfrak{b}_L)$ is a bijection. By the identity of Lemma 5 and the orthonormality of $\{e_k\}_{k=1}^M$, we have $\|\sum_{j=1}^M f_j(v)\phi(\widetilde{y}_j)\|_{\mathcal{H}_\mathcal{Y}} = \|\psi(v)\|$ for all $v \in \mathcal{V}$. The bijection between the output layers, the identity of Lemma 5 and this norm identity yield:

$$\{\mathbf{K}_M^{1/2} f : f \in \mathcal{F}_{\mathrm{FC}}\} = \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F}), \qquad \mathcal{F}_{\mathrm{NK}} = \Big\{ \sum_{k=1}^M \psi_k(\cdot) e_k : \psi \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F}) \Big\}. \tag{76}$$

By Lemma 5, the coordinates of $\widehat{\mu}_{\mathrm{NK}}(v)$ and $\mu^\star_{\widehat{\xi}}(v)$ in the orthonormal basis $\{e_k\}_{k=1}^M$ are $\widehat{\psi}(v) := \mathbf{K}_M^{1/2}\widehat{f}(v)$ and $\psi^\star(v) := \mathbf{K}_M^{1/2} f^\star(v)$. In these coordinates, that is, with the isometry $\iota(w) = \sum_{k=1}^M w_k e_k$, the loss (9) satisfies:

$$\widehat{\mathcal{L}}_{\mathrm{NK}}(\theta) = \mathbb{E}_n[\mathcal{L}_{\iota \circ \mathbf{K}_M^{1/2} f_\theta}(V, Z)] - \mathbb{E}_n[\|\widehat{\xi}(Z)\|^2_{\mathcal{H}_\mathcal{Y}}].$$

The last term on the right-hand side does not depend on θ. Consequently, by (76), $\widehat{\psi} \in \arg\min_{\psi \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})} \mathbb{E}_n[\mathcal{L}_{\iota\circ\psi}(V, Z)]$. Similarly, by the definition of $\mu^\star_{\widehat{\xi}}$ and (76), we have $\psi^\star \in \arg\min_{\psi \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})} \mathbb{E}[\mathcal{L}_{\iota\circ\psi}(V, Z) \mid \mathcal{D}_0]$. Thus, $\widehat{\psi}$ and $\psi^\star$ are the minimizers in (54) for the class $\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})$.

We now verify the two conditions of Lemma 8. Lemma 3 with $B = B_\mathcal{F}$ yields the metric entropy bound (70) for the class $\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})$ with $D_\mathcal{F} = c_{\mathrm{FC}} W L \log W$ and $A_\mathcal{F} = 4\sqrt{M}$. Thus, condition (i) is satisfied. Since Lemma 2 yields $\|\widehat{\xi}(z)\|_{\mathcal{H}_\mathcal{Y}} \le \sqrt{B_{\widehat{\xi}}}$ for all $z \in \mathcal{Z}$, condition (ii) is also satisfied.

We are now in a position to apply Lemma 8. With $B = B_\mathcal{F}$, $B' = \sqrt{B_{\widehat{\xi}}}$ and $\mathcal{F}^M = \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})$, we obtain:

$$\mathbb{E}\Big[\|\widehat{\mu}_{\mathrm{NK}} - \Pi_M \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} \,\Big|\, \mathcal{D}_0\Big] - c_3 \|\mu^\star_{\widehat{\xi}} - \Pi_M \widetilde{\mu}_{Y^1|V}\|^2_{L_2(P_V)} \lesssim \frac{W L \log W \log(Mn)}{n}.$$

Thus, after taking the expectation over $\mathcal{D}_0$, we obtain a bound for $\mathbf{A}$:

$$\mathbf{A} \lesssim \frac{W L \log W \log(Mn)}{n}.$$

Term B: The Approximation Error. As in the proof for Term B of the Deep Feature estimator in Appendix M, we apply Lemma 11, whose only nontrivial condition is $D \ge \bar{w} M$ where $D := \min_{1 \le i \le L-1} d_i$. Our choices of M and WL in Case 1 yield $WL/M^2 \asymp n^{a/(ab+a+2b)}$ and those in Case 2 yield $WL/M^2 \asymp n^{1/(b+1)} (\log n)^{-4d_y}$. Since L does not depend on n, we have $W/M^2 \to \infty$ as $n \to \infty$ in both cases. We then proceed as with the Deep Feature estimator to obtain $W \lesssim LD^2$ and $D/M \to \infty$ as $n \to \infty$. Consequently, Lemma 11 yields a network $\psi_c \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})$ such that, as in Appendix M,

$$\mathbb{E}\left[\Big\| \sum_{k=1}^M \psi_{c,k}(V) e_k - \Pi_M \mu_{Y^1|V}(V) \Big\|^2_{\mathcal{H}_\mathcal{Y}} \right] \lesssim \left(\frac{M}{DL}\right)^{2b} \lesssim \left(\frac{M^2}{WL}\right)^b.$$

By (76), the map $\sum_{k=1}^{M} \psi_{c,k}(\cdot)e_k$ belongs to $\mathcal{F}_{\mathrm{NK}}$. Consequently,

$$\mathbf{B} = \inf_{\mu \in \mathcal{F}_{\mathrm{NK}}} \mathbb{E}\big[\|\mu(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \leq \mathbb{E}\left[\Big\|\sum_{k=1}^{M} \psi_{c,k}(V)e_k - \Pi_M \mu_{Y^1|V}(V)\Big\|_{\mathcal{H}_\mathcal{Y}}^2\right] \lesssim \left(\frac{M^2}{WL}\right)^b.$$

Term C: The Nuisance Error. By the non-expansiveness of the projection operator, we have

$$\begin{aligned}\mathbf{C} &= \mathbb{E}\big[\|\Pi_M \widetilde{\mu}_{Y^1|V}(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \\ &\leq \mathbb{E}\big[\|\widetilde{\mu}_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big].\end{aligned}$$

Since the assumptions of the theorem include $\pi(x), \widehat{\pi}(x) \in [\varepsilon, 1)$ and $\sup_{x\in\mathcal{X}} \|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$, we can use Lemma 1 to bound the right-hand side of the last display:

$$\mathbf{C} \lesssim \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.$$

Term D: The Projection Error. As in the proof for Term $\mathbf{D}$ of the Deep Feature estimator in Appendix M, we apply Lemma 12. The only difference is that the grid points $\{\widetilde{y}_j\}_{j=1}^M$ of Assumption 8 directly satisfy the fill distance requirement $h \asymp M^{-1/d_y}$ of Lemma 12. Consequently, in Case 1 of the theorem,

$$\mathbf{D} = \mathbb{E}\big[\|\Pi_M \mu_{Y^1|V}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim M^{-2\tau/d_y}\, \mathbb{E}\Big[\|\mathfrak{p}^1(\cdot|V)\|_{L^2(\mathcal{Y})}^2\Big] \lesssim M^{-a}.$$

In Case 2 of the theorem, Lemma 12 yields the following bound for some $B_0 > 0$:

$$\mathbf{D} \lesssim \exp\Big(-B_0 M^{\frac{1}{2d_y}} \log M\Big). \tag{77}$$

Combining the bounds. In Case 1 of the theorem, we obtain the following inequality from (75) and the bounds of Terms $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$:

$$\begin{aligned}\mathbb{E}[\|\widehat{\mu}_{\mathrm{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] &\leq 3\mathbf{A} + c(\mathbf{B} + \mathbf{C} + \mathbf{D}) \\ &\lesssim \frac{WL \log W \log(Mn)}{n} + \left(\frac{M^2}{WL}\right)^b + M^{-a} \\ &\quad + \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.\end{aligned}$$

Recall our choices of $M \asymp n^{\frac{b}{ab+a+2b}}$ and $WL \asymp n^{\frac{a+2b}{ab+a+2b}}$. These choices yield $WL/n \asymp (M^2/(WL))^b \asymp M^{-a} \asymp n^{-\frac{ab}{ab+a+2b}}$. Therefore, we obtain the bound:

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim_{\log n} n^{-\frac{ab}{ab+a+2b}} + \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.$$

In Case 2, we instead obtain from (75), the bounds of Terms $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, and (77) that:

$$\begin{aligned}&\mathbb{E}[\|\widehat{\mu}_{\mathrm{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \\ &\lesssim \frac{WL \log W \log(Mn)}{n} + \left(\frac{M^2}{WL}\right)^b + \exp\Big(-B_0 M^{\frac{1}{2d_y}} \log M\Big) + \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.\end{aligned}$$

Recall our choices of $M \asymp (\log n)^{2d_y}$ and $WL \asymp n^{\frac{1}{b+1}}$. As in the proof of Theorem 3, we have $\exp(-B_0 M^{\frac{1}{2d_y}} \log M) \leq n^{-1}$ for all sufficiently large n. These choices also yield $\frac{WL \log W \log(Mn)}{n} + (M^2/(WL))^b \lesssim_{\log n} n^{-\frac{b}{b+1}}$. Therefore, we obtain the bound:

$$\mathbb{E}[\|\widehat{\mu}_{\mathrm{NK}}(V) - \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2] \lesssim_{\log n} n^{-\frac{b}{b+1}} + \min\big\{\mathcal{R}_\pi^2(\widehat{\pi}), \mathcal{R}_{\mu_0}^2(\widehat{\mu}_0)\big\}.$$

■

O Miscellaneous Lemmas

Lemma 9 (Lemma 1, restated)**.** *Let Assumptions 1 and 2 hold. In addition, assume that* $\widehat{\pi}(x), \pi(x) \in [\varepsilon, 1)$ *for all* $x \in \mathcal{X}$*, and that* $\sup_{x\in\mathcal{X}} \|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$ *for some constant* $B_{\mu_0} > 0$*. Then we have the following bound:*

$$\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \lesssim \min\big\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\big\}.$$

Proof. By the identification result (2), the independence of Z and $\mathcal{D}_0$ and the definitions of $\widehat{\xi}$ and ξ,

$$\begin{aligned}
\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V) &= \mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mathbb{E}[\xi(Z)|V] \\
&= \mathbb{E}[\widehat{\xi}(Z) - \xi(Z)|V, \mathcal{D}_0] \\
&= \mathbb{E}\left[\left(\frac{A}{\widehat{\pi}(X)} - \frac{A}{\pi(X)}\right)(\phi(Y) - \mu_0(X))\middle|V, \mathcal{D}_0\right] \\
&\quad - \mathbb{E}\left[\left(\frac{A}{\widehat{\pi}(X)} - 1\right)(\widehat{\mu}_0(X) - \mu_0(X))\middle|V, \mathcal{D}_0\right].
\end{aligned}$$

Since $V = \eta(X)$ is a function of X and $\mathbb{E}[A\phi(Y)|X] = \pi(X)\mathbb{E}[\phi(Y)|X, A = 1] = \pi(X)\mu_0(X) = \mathbb{E}[A\mu_0(X)|X]$, the law of iterated expectations, conditionally on $(X, \mathcal{D}_0)$, yields that the first term on the right-hand side is zero. Thus, using $\mathbb{E}[A|X] = \pi(X)$, $\widehat{\pi}(X) \geq \varepsilon$, Jensen's inequality and the Cauchy-Schwarz inequality, we obtain:

$$\begin{aligned}
&\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \\
&\lesssim \mathbb{E}\left[\mathbb{E}\Big[|\widehat{\pi}(X) - \pi(X)| \cdot \|\widehat{\mu}_0(X) - \mu_0(X)\|_{\mathcal{H}_\mathcal{Y}}\Big|V, \mathcal{D}_0\Big]^2\right] \\
&\leq \mathbb{E}\Big[\mathbb{E}\Big[|\widehat{\pi}(X) - \pi(X)|^2\Big|V, \mathcal{D}_0\Big]\mathbb{E}\Big[\|\widehat{\mu}_0(X) - \mu_0(X)\|^2_{\mathcal{H}_\mathcal{Y}}\Big|V, \mathcal{D}_0\Big]\Big]. \qquad (78)
\end{aligned}$$

We now bound the first factor on the right-hand side of (78) by a constant. Since both $\pi(X) < 1$ and $\widehat{\pi}(X) < 1$, this factor is at most 1. Thus, (78) yields:

$$\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \lesssim \mathbb{E}\Big[\mathbb{E}\Big[\|\widehat{\mu}_0(X) - \mu_0(X)\|^2_{\mathcal{H}_\mathcal{Y}}\Big|V, \mathcal{D}_0\Big]\Big] = \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0). \quad (79)$$

Alternatively, we bound the second factor on the right-hand side of (78) by a constant. Let Y' be a copy of Y that is independent of Y given X and $A = 1$. Then,

$$\begin{aligned}
\|\mu_0(X)\|^2_{\mathcal{H}_\mathcal{Y}} &= \langle \mu_0(X), \mu_0(X)\rangle_{\mathcal{H}_\mathcal{Y}} \\
&= \langle \mathbb{E}[\phi(Y) \mid X, A = 1], \mathbb{E}[\phi(Y') \mid X, A = 1]\rangle_{\mathcal{H}_\mathcal{Y}} \\
&= \mathbb{E}[\langle \phi(Y), \phi(Y')\rangle_{\mathcal{H}_\mathcal{Y}} \mid X, A = 1] = \mathbb{E}[k_\mathcal{Y}(Y, Y') \mid X, A = 1] \leq B_k.
\end{aligned}$$

Moreover, by assumption, $\|\widehat{\mu}_0(x)\|_{\mathcal{H}_\mathcal{Y}} \leq B_{\mu_0}$ for all x. Consequently, $\mathbb{E}[\|\widehat{\mu}_0(X) - \mu_0(X)\|^2_{\mathcal{H}_\mathcal{Y}}|V, \mathcal{D}_0] \leq (B_{\mu_0} + \sqrt{B_k})^2$ almost surely. Therefore, (78) yields:

$$\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \lesssim \mathbb{E}\Big[\mathbb{E}\Big[|\widehat{\pi}(X) - \pi(X)|^2\Big|V, \mathcal{D}_0\Big]\Big] = \mathcal{R}^2_\pi(\widehat{\pi}). \quad (80)$$

Combining (79) and (80) yields:

$$\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \lesssim \min\big\{\mathcal{R}^2_\pi(\widehat{\pi}), \mathcal{R}^2_{\mu_0}(\widehat{\mu}_0)\big\}.$$

□

Proof of (11). We also derive the product bound (11) in Remark 1. We apply the Cauchy-Schwarz inequality for the outer expectation on the right-hand side of (78), followed by Jensen's inequality:

$$\mathbb{E}\Big[\|\mathbb{E}[\widehat{\xi}(Z)|V, \mathcal{D}_0] - \mu_{Y^1|V}(V)\|^2_{\mathcal{H}_\mathcal{Y}}\Big] \lesssim \mathbb{E}\big[(\widehat{\pi}(X) - \pi(X))^4\big]^{1/2}\mathbb{E}\big[\|\widehat{\mu}_0(X) - \mu_0(X)\|^4_{\mathcal{H}_\mathcal{Y}}\big]^{1/2}.$$

□

Lemma 10 (Sobolev Regularity of Projection Coefficients)**.** *Let Assumptions 2, 4 and 6 hold. Let $\Pi_M \colon \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ be the orthogonal projection onto a subspace $\mathcal{H}_M \subseteq \mathcal{H}_\mathcal{Y}$ with an orthonormal basis $\{e_k\}_{k=1}^M$, and denote $c_k(v) \coloneqq \int_\mathcal{Y} e_k(y)\mathfrak{p}^1(y|v)\,\mathrm{d}y$ for $v \in \mathcal{V}$, so that $\Pi_M \mu_{Y^1|V}(v) = \sum_{k=1}^M c_k(v)e_k$. Denote by $|\mathcal{Y}|$ the Lebesgue measure of $\mathcal{Y}$ and by N_r the number of multi-indices $\alpha \in \mathbb{N}_0^{d_v}$ with $|\alpha| \leq r$. Then, the following inequality holds:*

$$\sum_{k=1}^M \|c_k\|^2_{W^{r,2}(\mathcal{V})} \leq N_r B_k |\mathcal{Y}| \int_\mathcal{Y} \|\mathfrak{p}^1(y|\cdot)\|^2_{W^{r,2}(\mathcal{V})}\,\mathrm{d}y.$$

Proof. Let $q \in L^2(\mathcal{Y})$, and denote $\mu_q \coloneqq \int_\mathcal{Y} \phi(y)q(y)\,\mathrm{d}y \in \mathcal{H}_\mathcal{Y}$. By the reproducing property, $\langle \mu_q, h\rangle_{\mathcal{H}_\mathcal{Y}} = \int_\mathcal{Y} h(y)q(y)\,\mathrm{d}y$ for all $h \in \mathcal{H}_\mathcal{Y}$. By Assumption 2 and the Cauchy-Schwarz inequality, we have:

$$\|\mu_q\|^2_{\mathcal{H}_\mathcal{Y}} = \int_\mathcal{Y}\int_\mathcal{Y} k_\mathcal{Y}(y,y')q(y)q(y')\,\mathrm{d}y\,\mathrm{d}y' \leq B_k\left(\int_\mathcal{Y} |q(y)|\,\mathrm{d}y\right)^2 \leq B_k|\mathcal{Y}|\|q\|^2_{L^2(\mathcal{Y})}.$$

Thus, Bessel's inequality for the orthonormal set $\{e_k\}_{k=1}^M$ and the last display yield:

$$\sum_{k=1}^M \left(\int_\mathcal{Y} e_k(y)q(y)\,\mathrm{d}y\right)^2 = \sum_{k=1}^M \langle \mu_q, e_k\rangle^2_{\mathcal{H}_\mathcal{Y}} \leq \|\mu_q\|^2_{\mathcal{H}_\mathcal{Y}} \leq B_k|\mathcal{Y}|\|q\|^2_{L^2(\mathcal{Y})}. \tag{81}$$

Next, we apply (81) to the coefficients c_k. Since $\mathfrak{p}^1(\cdot|v)$ is the conditional density of Y^1 given $V = v$, the CCME can be written as the following Bochner integral:

$$\mu_{Y^1|V}(v) = \mathbb{E}[\phi(Y^1)|V = v] = \int_\mathcal{Y} \phi(y)\mathfrak{p}^1(y|v)\,\mathrm{d}y. \tag{82}$$

Thus, $\mu_{Y^1|V}(v) = \mu_{\mathfrak{p}^1(\cdot|v)}$. By the reproducing property, the coefficients of $\Pi_M\mu_{Y^1|V}(v)$ are then $\langle \mu_{Y^1|V}(v), e_k\rangle_{\mathcal{H}_\mathcal{Y}} = \int_\mathcal{Y} e_k(y)\mathfrak{p}^1(y|v)\,\mathrm{d}y = c_k(v)$. For a multi-index α with $|\alpha| \leq r$, we apply the derivative operator D_v^α. Under Assumption 6, and since $|e_k(y)| \leq \sqrt{B_k}$, we can differentiate under the integral sign:

$$D_v^\alpha c_k(v) = \int_\mathcal{Y} e_k(y)\, D_v^\alpha \mathfrak{p}^1(y|v)\,\mathrm{d}y.$$

Thus, applying (81) with $q = D_v^\alpha \mathfrak{p}^1(\cdot|v)$, integrating over $v \in \mathcal{V}$, summing over $|\alpha| \leq r$, and applying Fubini's theorem and the Cauchy-Schwarz inequality to the sum over α yields:

$$\begin{aligned}\sum_{k=1}^M \|c_k\|^2_{W^{r,2}(\mathcal{V})} &\leq N_r B_k|\mathcal{Y}| \int_\mathcal{V} \sum_{|\alpha|\leq r} \int_\mathcal{Y} \left|D_v^\alpha \mathfrak{p}^1(y|v)\right|^2 \mathrm{d}y\,\mathrm{d}v \\ &\leq N_r B_k|\mathcal{Y}| \int_\mathcal{Y} \|\mathfrak{p}^1(y|\cdot)\|^2_{W^{r,2}(\mathcal{V})}\,\mathrm{d}y.\end{aligned}$$

□

Lemma 11 (Approximation by a Fully Connected Network)**.** *Let Assumptions 2, 4 and 6 hold, and denote $b \coloneqq 2r/d_v$. Let $\Pi_M \colon \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ be the orthogonal projection onto a subspace $\mathcal{H}_M \subseteq \mathcal{H}_\mathcal{Y}$ with an orthonormal basis $\{e_k\}_{k=1}^M$. Then, there exist constants L_0 and $\bar{w} \geq 2$ that only depend on r and d_v, a constant $\bar{B} > 0$ that only depends on $\mathcal{V}$, $\mathcal{Y}$, r, B_k and $\mathfrak{p}^1$, and a constant $C > 0$ that does not depend on M or $\mathcal{H}_M$, such that, for any $B_\mathcal{F}^2 \geq \bar{B}$, any $L \geq \max\{L_0, 3\}$ and any widths $\mathbf{d} = (d_v, d_1, \ldots, d_{L-1}, M)$ of the networks (12) whose smallest hidden width $D \coloneqq \min_{1\leq i\leq L-1} d_i$ satisfies $D \geq \bar{w}M$, there is $\psi_c \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_\mathcal{F})$ with*

$$\mathbb{E}\left[\left\|\sum_{k=1}^M \psi_{c,k}(V)e_k - \Pi_M\mu_{Y^1|V}(V)\right\|^2_{\mathcal{H}_\mathcal{Y}}\right] \leq C\left(\frac{M}{DL}\right)^{2b}.$$

Proof. With c_k as in Lemma 10, we have $\Pi_M\mu_{Y^1|V}(v) = \sum_{k=1}^M c_k(v)e_k$. Our main idea of the proof is to take the networks of [Yan25, Theorem 2.1] that approximate the coefficients c_k, to clip

their outputs, and to place these networks in parallel in one network of $\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_{\mathcal{F}})$. In this proof, the constants in $\lesssim$ do not depend on M, $\mathcal{H}_M$ or the widths $d_1, \dots, d_{L-1}$.

We first extend the coefficients to a cube, on which [Yan25, Theorem 2.1] holds. Fix a cube $\mathcal{Q} \supseteq \mathcal{V}$. According to [Ste70, Chapter VI, Theorem 5], the condition that $\mathcal{V}$ is bounded with a Lipschitz boundary (Assumption 4) yields a linear operator that extends every function on $\mathcal{V}$ to a function on $\mathbb{R}^{d_v}$ and maps $W^{r,2}(\mathcal{V})$ continuously into $W^{r,2}(\mathbb{R}^{d_v})$. We denote by $\bar{c}_k$ the image of c_k under this operator. Thus, $\bar{c}_k = c_k$ on $\mathcal{V}$. Moreover, there is a constant $C_{\mathcal{V}}$ that only depends on $\mathcal{V}$ and r such that $\|\bar{c}_k\|_{W^{r,2}(\mathcal{Q})} \leq C_{\mathcal{V}} \|c_k\|_{W^{r,2}(\mathcal{V})}$ for every k. This bound, Lemma 10 and Assumption 6 then yield:

$$\sum_{k=1}^{M} \|\bar{c}_k\|^2_{W^{r,2}(\mathcal{Q})} \leq C_{\mathcal{V}}^2 N_r B_k |\mathcal{Y}| \int_{\mathcal{Y}} \|\mathfrak{p}^1(y|\cdot)\|^2_{W^{r,2}(\mathcal{V})} \, \mathrm{d}y < \infty. \tag{83}$$

Since $r > d_v/2$ by Assumption 6, the space $W^{r,2}(\mathcal{Q})$ embeds continuously into the continuous functions on $\overline{\mathcal{Q}}$. Thus, there is a constant $C_{\mathcal{Q}}$ such that $s_k := \sup_{\mathcal{Q}} |\bar{c}_k| \leq C_{\mathcal{Q}} \|\bar{c}_k\|_{W^{r,2}(\mathcal{Q})}$ for every k.

We now construct ψ_c. Let L_0 and $\bar{w} \geq 2$ be integers that only depend on r and d_v such that [Yan25, Theorem 2.1] provides its approximation bound for the networks (12) with one output, depth at least $L_0 - 1$ and width at least $\bar{w}$ in every hidden layer, and denote $w := \lfloor D/M \rfloor$. Since $\bar{w} \leq D/M$ and $\bar{w}$ is an integer, we have $\bar{w} \leq \lfloor D/M \rfloor = w$. In addition, we have from our assumption that $L \geq L_0$. Therefore, according to [Yan25, Theorem 2.1], for every k, there exists a network $\widetilde{\psi}_k$ of the form (12) with one output, depth $L - 1$ and width w in every hidden layer such that:

$$\|\widetilde{\psi}_k - \bar{c}_k\|^2_{L^2(\mathcal{Q})} \lesssim (wL)^{-2b} \|\bar{c}_k\|^2_{W^{r,2}(\mathcal{Q})}. \tag{84}$$

Since (84) only bounds the $L^2(\mathcal{Q})$ error, the values of $\widetilde{\psi}_k$ need not lie in $[-s_k, s_k]$, while the class $\mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_{\mathcal{F}})$ requires a bound on $\sup_{v \in \mathcal{V}} \|\psi_c(v)\|$. Thus, we replace the linear output unit in $\widetilde{\psi}_k$ by the two ReLU units $\sigma_{\mathrm{relu}}(\widetilde{\psi}_k(v) + s_k)$ and $\sigma_{\mathrm{relu}}(\widetilde{\psi}_k(v) - s_k)$. Then, we add a linear output unit with the following value:

$$\sigma_{\mathrm{relu}}(\widetilde{\psi}_k(v) + s_k) - \sigma_{\mathrm{relu}}(\widetilde{\psi}_k(v) - s_k) - s_k = \min\{\max\{\widetilde{\psi}_k(v), -s_k\}, s_k\}.$$

The resulting network has $L - 2$ hidden layers of width w, a hidden layer of width 2 and a linear output. Since $D \geq \bar{w} M \geq 2M$, we can place the resulting networks for $k = 1, \dots, M$ in parallel in one network ψ_c with the widths $\mathbf{d}$.

Finally, we bound ψ_c. By construction, $\psi_{c,k}(v) = \min\{\max\{\widetilde{\psi}_k(v), -s_k\}, s_k\}$ for every k. Since $|\psi_{c,k}| \leq s_k$, the bound on s_k and (83) yield:

$$\sup_{v \in \mathcal{V}} \|\psi_c(v)\|^2 \leq \sum_{k=1}^{M} s_k^2 \leq C_{\mathcal{Q}}^2 C_{\mathcal{V}}^2 N_r B_k |\mathcal{Y}| \int_{\mathcal{Y}} \|\mathfrak{p}^1(y|\cdot)\|^2_{W^{r,2}(\mathcal{V})} \, \mathrm{d}y =: \bar{B}. \tag{85}$$

Since $\mathcal{Q}$ depends only on $\mathcal{V}$, the constant $\bar{B}$ depends only on $\mathcal{V}$, $\mathcal{Y}$, r, B_k and $\mathfrak{p}^1$. Thus, $\psi_c \in \mathcal{F}_{\mathrm{NN}}(\mathbf{d}, B_{\mathcal{F}})$ whenever $B_{\mathcal{F}}^2 \geq \bar{B}$. Moreover, for every $v \in \mathcal{V}$, we have $c_k(v) = \bar{c}_k(v) \in [-s_k, s_k]$. Since the map $t \mapsto \min\{\max\{t, -s_k\}, s_k\}$ is 1-Lipschitz and fixes every point of $[-s_k, s_k]$, we obtain $|\psi_{c,k}(v) - c_k(v)| \leq \left|\widetilde{\psi}_k(v) - c_k(v)\right|$ for $v \in \mathcal{V}$. Thus, the orthonormality of $\{e_k\}_{k=1}^M$, the bound $\sup_{v \in \mathcal{V}} p_V(v) < \infty$, (84), (83) and $w = \lfloor D/M \rfloor \geq D/(2M)$, respectively, yield the bound in the lemma:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^{M} \psi_{c,k}(V) e_k - \Pi_M \mu_{Y^1|V}(V)\right\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] &= \sum_{k=1}^{M} \mathbb{E}[(\psi_{c,k}(V) - c_k(V))^2] \\
&\leq \sum_{k=1}^{M} \mathbb{E}[(\widetilde{\psi}_k(V) - c_k(V))^2] \\
&\lesssim (wL)^{-2b} \sum_{k=1}^{M} \|\bar{c}_k\|^2_{W^{r,2}(\mathcal{Q})} \lesssim \left(\frac{M}{DL}\right)^{2b}.
\end{aligned}$$

□

Lemma 12 (Projection Error Bound). *Let $\mathcal{Y} \subset \mathbb{R}^{d_y}$ be a bounded domain with a Lipschitz boundary. Let $\{\widetilde{y}_j\}_{j=1}^M \subset \mathcal{Y}$ be a set of grid points with fill distance $h \coloneqq \sup_{y\in\mathcal{Y}} \min_j \|y - \widetilde{y}_j\| \asymp M^{-1/d_y}$. Let $\Pi_M : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ denote the orthogonal projection from $\mathcal{H}_\mathcal{Y}$ onto the subspace spanned by $\{k_\mathcal{Y}(\cdot, \widetilde{y}_j)\}_{j=1}^M = \{\phi(\widetilde{y}_j)\}_{j=1}^M$. In the following bounds, the constants in $\lesssim$ do not depend on M or the grid points.*

1. ***(Sobolev Kernels)*** *Assume $k_\mathcal{Y}(y, y') = \varphi(y - y')$ for all $y, y' \in \mathcal{Y}$, where $\widehat{\varphi}(\omega) \lesssim (1 + \|\omega\|^2)^{-\tau}$ for some $\tau > d_y/2$. Suppose that $\mathfrak{p}^1(\cdot|V) \in L^2(\mathcal{Y})$ almost surely. Then, the expected approximation error satisfies:*

$$\mathbb{E}\big[\|\mu_{Y^1|V}(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim M^{-\frac{2\tau}{d_y}} \mathbb{E}\Big[\|\mathfrak{p}^1(\cdot|V)\|_{L^2(\mathcal{Y})}^2\Big]. \tag{86}$$

2. ***(Gaussian Kernels)*** *Assume that $\mathcal{Y}$ satisfies a uniform interior cone condition (Definition 7) and that $\mathcal{H}_\mathcal{Y}$ is the RKHS generated by a Gaussian kernel. Suppose that $\mathfrak{p}^1(\cdot|V) \in L^2(\mathcal{Y})$ almost surely. Then, there exists a constant $B_0 > 0$ such that:*

$$\mathbb{E}\big[\|\mu_{Y^1|V}(V) - \Pi_M \mu_{Y^1|V}(V)\|_{\mathcal{H}_\mathcal{Y}}^2\big] \lesssim \exp\Big(-B_0 M^{\frac{1}{2d_y}} \log M\Big) \mathbb{E}\Big[\|\mathfrak{p}^1(\cdot|V)\|_{L^2(\mathcal{Y})}^2\Big]. \tag{87}$$

Proof. Given any $v \in \mathcal{V}$, we denote $\mu_v \coloneqq \mu_{Y^1|V}(v)$ and $e_v \coloneqq \mu_v - \Pi_M \mu_v$. We first show, for both kernels, that e_v vanishes at the grid points and that $\|e_v\|_{\mathcal{H}_\mathcal{Y}}^2 \le \|e_v\|_{L^2(\mathcal{Y})} \|\mathfrak{p}^1(\cdot|v)\|_{L^2(\mathcal{Y})}$.

Since e_v is orthogonal to $\operatorname{span}\{\phi(\widetilde{y}_j)\}_{j=1}^M$, which contains $\Pi_M \mu_v$, we have:

$$\|e_v\|_{\mathcal{H}_\mathcal{Y}}^2 = \langle e_v, \mu_v - \Pi_M \mu_v \rangle_{\mathcal{H}_\mathcal{Y}} = \langle e_v, \mu_v \rangle_{\mathcal{H}_\mathcal{Y}}.$$

Since e_v is orthogonal to $\phi(\widetilde{y}_j)$, the reproducing property also yields $e_v(\widetilde{y}_j) = \langle e_v, \phi(\widetilde{y}_j) \rangle_{\mathcal{H}_\mathcal{Y}} = 0$ for all $j = 1, \dots, M$. Moreover, by (82) and the reproducing property, we obtain:

$$\|e_v\|_{\mathcal{H}_\mathcal{Y}}^2 = \left\langle e_v, \int_\mathcal{Y} \phi(y) \mathfrak{p}^1(y|v)\, \mathrm{d}y \right\rangle_{\mathcal{H}_\mathcal{Y}} = \int_\mathcal{Y} e_v(y) \mathfrak{p}^1(y|v)\, \mathrm{d}y = \langle e_v, \mathfrak{p}^1(\cdot|v) \rangle_{L^2(\mathcal{Y})}.$$

Thus, applying the Cauchy-Schwarz inequality in $L^2(\mathcal{Y})$, we obtain:

$$\|e_v\|_{\mathcal{H}_\mathcal{Y}}^2 \le \|e_v\|_{L^2(\mathcal{Y})} \|\mathfrak{p}^1(\cdot|v)\|_{L^2(\mathcal{Y})}. \tag{88}$$

We now prove (86). To bound $\|e_v\|_{L^2(\mathcal{Y})}$, we invoke the sampling inequality from [ALdST07, Theorem 4.1] for Sobolev spaces of order $\tau > d_y/2$ on bounded connected open sets with a Lipschitz boundary. Since $\mathcal{Y}$ is bounded with a Lipschitz boundary, $\mathcal{Y}$ has finitely many connected components $\mathcal{Y}_1, \dots, \mathcal{Y}_J$ with Lipschitz boundaries, whose closures are pairwise disjoint. Thus, we can apply the sampling inequality on each $\mathcal{Y}_l$. More precisely, let h_l be the fill distance of the grid points in $\mathcal{Y}_l$. According to [ALdST07, Theorem 4.1], for each l, there is a threshold $h_{0,l} > 0$, which depends only on $\mathcal{Y}_l$ and τ, such that, whenever $h_l \le h_{0,l}$, every $f \in W^{\tau,2}(\mathcal{Y}_l)$ that vanishes at the grid points in $\mathcal{Y}_l$ satisfies:

$$\|f\|_{L^2(\mathcal{Y}_l)} \lesssim h_l^\tau \|f\|_{W^{\tau,2}(\mathcal{Y}_l)}. \tag{89}$$

Since the sampling inequality applies to each connected component with only the grid points in that component, we need to make sure that the fill distance h is smaller than the distances between the connected components. To this end, let $\delta > 0$ be smaller than the distance between the closures of any two different components, and let $h_1 \coloneqq \min\{\delta, h_{0,1}, \dots, h_{0,J}\}$. Choose a sufficiently large M so that $h \le h_1$. Since $h \le \delta$, every grid point within distance h of a point $y \in \mathcal{Y}_l$ lies in $\mathcal{Y}_l$. Thus, by the definition of the fill distance, $h_l \le h$ for every l. Consequently, $h_l \le h_{0,l}$ for every l. Thus, since e_v vanishes at the grid points, for such h, the sampling inequality (89) with $f = e_v$ yields, for every l:

$$\|e_v\|_{L^2(\mathcal{Y}_l)} \lesssim h_l^\tau \|e_v\|_{W^{\tau,2}(\mathcal{Y}_l)} \le h^\tau \|e_v\|_{W^{\tau,2}(\mathcal{Y}_l)}.$$

Moreover, by Proposition 3 and the bound $\widehat{\varphi}(\omega) \lesssim (1 + \|\omega\|^2)^{-\tau}$, every $F \in \mathcal{H}_{\mathbb{R}^{d_y}}$ satisfies $\|F\|_{W^{\tau,2}(\mathbb{R}^{d_y})}^2 \lesssim \int_{\mathbb{R}^{d_y}} \big|\widehat{F}(\omega)\big|^2 / \widehat{\varphi}(\omega)\, \mathrm{d}\omega = \|F\|_{\mathcal{H}_{\mathbb{R}^{d_y}}}^2$. Taking the infimum over the extensions F of e_v (Proposition 3) yields $\|e_v\|_{W^{\tau,2}(\mathcal{Y})} \lesssim \|e_v\|_{\mathcal{H}_\mathcal{Y}}$. Summing the squares of the inequalities

$\|e_v\|_{L^2(\mathcal{Y}_l)} \lesssim h^\tau \|e_v\|_{W^{\tau,2}(\mathcal{Y}_l)}$ over $l = 1, \dots, J$, and using $\|e_v\|_{W^{\tau,2}(\mathcal{Y}_l)} \leq \|e_v\|_{W^{\tau,2}(\mathcal{Y})}$ and this bound, we obtain:

$$\|e_v\|_{L^2(\mathcal{Y})} = \Big(\sum_{l=1}^{J} \|e_v\|^2_{L^2(\mathcal{Y}_l)}\Big)^{1/2} \lesssim h^\tau \Big(\sum_{l=1}^{J} \|e_v\|^2_{W^{\tau,2}(\mathcal{Y}_l)}\Big)^{1/2} \lesssim h^\tau \|e_v\|_{W^{\tau,2}(\mathcal{Y})} \lesssim h^\tau \|e_v\|_{\mathcal{H}_\mathcal{Y}}.$$

Plugging this bound back into the right-hand side of (88) yields:

$$\|e_v\|^2_{\mathcal{H}_\mathcal{Y}} \lesssim h^\tau \|e_v\|_{\mathcal{H}_\mathcal{Y}} \|\mathfrak{p}^1(\cdot|v)\|_{L^2(\mathcal{Y})}.$$

Dividing both sides by $\|e_v\|_{\mathcal{H}_\mathcal{Y}}$, we get:

$$\|e_v\|_{\mathcal{H}_\mathcal{Y}} \lesssim h^\tau \|\mathfrak{p}^1(\cdot|v)\|_{L^2(\mathcal{Y})}.$$

Squaring both sides and taking the expectation with respect to V yields:

$$\mathbb{E}\big[\|e_V\|^2_{\mathcal{H}_\mathcal{Y}}\big] \lesssim h^{2\tau}\, \mathbb{E}\Big[\|\mathfrak{p}^1(\cdot|V)\|^2_{L^2(\mathcal{Y})}\Big].$$

Thus, we obtain (86) by substituting the fill distance $h \asymp M^{-1/d_y}$ for every M with $h \leq h_1$.

We now prove (87). In this case, $\mathcal{H}_\mathcal{Y}$ is an RKHS with a Gaussian kernel. To bound the L^2-norm of e_v, we invoke the sampling inequality of [RZ10, Theorem 3.5] for bounded domains with a Lipschitz boundary that satisfy an interior cone condition, which states that there exist constants $B_1, h_0 > 0$, where B_1 depends only on $\mathcal{Y}$, such that, whenever the fill distance satisfies $h \leq h_0$, every $u \in \mathcal{H}_\mathcal{Y}$ satisfies:

$$\|u\|_{L^2(\mathcal{Y})} \leq \exp\Big(\frac{B_1 \log h}{\sqrt{h}}\Big)\|u\|_{\mathcal{H}_\mathcal{Y}} + \sup_{1\leq j\leq M} |u(\widetilde{y}_j)|. \tag{90}$$

With $u = e_v$ in (90) and $e_v(\widetilde{y}_j) = 0$ for all j, we obtain:

$$\|e_v\|_{L^2(\mathcal{Y})} \leq \exp\Big(\frac{B_1 \log h}{\sqrt{h}}\Big)\|e_v\|_{\mathcal{H}_\mathcal{Y}}. \tag{91}$$

We then substitute (91) back into the right-hand side of (88):

$$\|e_v\|^2_{\mathcal{H}_\mathcal{Y}} \leq \Big[\exp\Big(\frac{B_1 \log h}{\sqrt{h}}\Big)\|e_v\|_{\mathcal{H}_\mathcal{Y}}\Big]\|\mathfrak{p}^1(\cdot|v)\|_{L^2(\mathcal{Y})}.$$

Dividing both sides by $\|e_v\|_{\mathcal{H}_\mathcal{Y}}$ yields:

$$\|e_v\|_{\mathcal{H}_\mathcal{Y}} \leq \exp\Big(\frac{B_1 \log h}{\sqrt{h}}\Big)\|\mathfrak{p}^1(\cdot|v)\|_{L^2(\mathcal{Y})}.$$

Squaring both sides and taking the expectation with respect to V, we obtain:

$$\mathbb{E}\big[\|e_V\|^2_{\mathcal{H}_\mathcal{Y}}\big] \leq \exp\Big(\frac{2B_1 \log h}{\sqrt{h}}\Big)\mathbb{E}\big[\|\mathfrak{p}^1(\cdot|V)\|^2_{L^2(\mathcal{Y})}\big].$$

Let $c_2 > 0$ be a constant with $h \leq c_2 M^{-1/d_y}$. For every $M \geq c_2^{2d_y}$, we have $c_2 M^{-1/d_y} \leq 1$ and $\log c_2 \leq \log M/(2d_y)$. Since $h \mapsto \log h/\sqrt{h}$ is increasing on $(0, 1]$, these two inequalities yield:

$$\frac{2B_1 \log h}{\sqrt{h}} \leq \frac{2B_1 \log(c_2 M^{-1/d_y})}{\sqrt{c_2 M^{-1/d_y}}} = \frac{2B_1}{\sqrt{c_2}}\Big(\log c_2 - \frac{\log M}{d_y}\Big) M^{\frac{1}{2d_y}} \leq -\frac{B_1}{d_y\sqrt{c_2}} M^{\frac{1}{2d_y}} \log M.$$

Thus, (87) holds with $B_0 := B_1/(d_y\sqrt{c_2})$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction explain the paper's contributions and scope: the proposed Conditional Counterfactual Mean Embeddings (CCME) framework, the three practical estimators, our theoretical analysis (finite-sample convergence rates), and the experimental results of these estimators.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss a limitation of our work in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full set of assumptions is stated in the main section. The complete proofs for all theoretical results are provided in the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the information needed to reproduce the main experimental results, including hyperparameters, setups, and data-generating processes, is detailed in the main section and appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code, with instructions to reproduce every experiment, is available at https://github.com/donlap/Conditional-Counterfactual-Mean-Embeddings.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All details about experimental settings are explained in the main section. Complete details, such as hyperparameters and optimizers, are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistical significance correctly. This includes standard errors and pointwise 50th and 90th percentile intervals, as detailed in the Experiments section and Figures of the main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computer resources required to reproduce the experiments is detailed in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our paper focuses on theoretical analysis. This research does not involve human participants, sensitive personal data, or applications with negative social impacts.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper focuses on foundational theoretical research. Our experiments are also limited to synthetic and semi-synthetic data; the work does not have direct, foreseeable societal impacts or specific real-world deployments; therefore, it does not include a broader impact discussion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This research focuses on theoretical analysis and utilizes a semi-synthetic MNIST dataset for evaluation. It does not involve the development or release of high-risk models or scraped datasets; therefore, specific safeguards are not required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We use no existing code or pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code at `https://github.com/donlap/Conditional-Counterfactual-Mean-Embeddings` under the MIT license, with a README that documents installation and how to run each experiment.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve any crowdsourcing, human subjects research, or direct data collection from participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve any human subjects research; therefore, Institutional Review Board (IRB) approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The use of LLMs in this paper was strictly limited to writing, editing, and formatting assistance, which does not require a declaration under the current policy.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.