Density Ratio-based Proxy Causal Learning Without Density Ratios

Bariscan Bozkurt
Gatsby Computational Neuroscience Unit

Ben Deaner
University College London

Dimitri Meunier
Gatsby Computational Neuroscience Unit

Liyuan Xu
Gatsby Computational Neuroscience Unit

Arthur Gretton
Gatsby Computational Neuroscience Unit

Abstract

We address the setting of Proxy Causal Learning (PCL), which has the goal of estimating causal effects from observed data in the presence of hidden confounding. Proxy methods accomplish this task using two proxy variables related to the latent confounder: a treatment proxy (related to the treatment) and an outcome proxy (related to the outcome). Two approaches have been proposed to perform causal effect estimation given proxy variables; however only one of these has found mainstream acceptance, since the other was understood to require density ratio estimation - a challenging task in high dimensions. In the present work, we propose a practical and effective implementation of the second approach, which bypasses explicit density ratio estimation and is suitable for continuous and high-dimensional treatments. We employ kernel ridge regression to derive estimators, resulting in simple closed-form solutions for dose-response and conditional dose-response curves, along with consistency guarantees. Our methods empirically demonstrate superior or comparable performance to existing frameworks on synthetic and real-world datasets.

1 INTRODUCTION

Causal inference aims to measure the impact of interventions on real-world outcomes, a crucial task across various scientific disciplines. Examples include assessing how changes in flight ticket prices will alter consumer demand (Blundell et al., 2012), the consequences of grade retention on students' cognitive development (Fruehwirth et al., 2016), the impact of medical treatments on patient health (Connors et al., 1996; Choi et al., 2002), and evaluating policies such as Job Corps (Schochet et al., 2008). In this context, the intervention is referred to as a *treatment*, which affects the *outcome*. However, estimating the causal relationship between treatment and outcome is a challenging task due to confounding variables - factors that influence both the treatment and the outcome - potentially leading to spurious correlations.

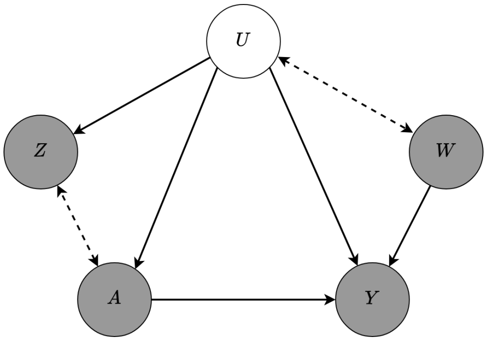


Figure 1: An instance of a Directed Acyclic Graph (DAG) for the PCL setting, which satisfies the required Assumption (3.2). In this graph, the gray circles denote the observed variables: A denotes the treatment, Y denotes the outcome, Z denotes the treatment proxy, and W denotes the outcome proxy. The white circle denotes the unobserved confounding variable U. Bi-directional dotted arrows indicate that either direction in the DAG is possible, or that both variables may share a common ancestor.

One widely used assumption is that no unobserved

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

confounding variable exists (Pearl and Robins, 1995), which allows causal effect to be estimated via regression or backdoor adjustment (Pearl, 2009). Although there are numerous methods relying on this assumption, e.g. (Hill, 2011; Johansson et al., 2016; Yao et al., 2018; Singh et al., 2023), it is often restrictive since it requires measuring all the covariates that account for the confounding variables. Recently, a growing literature has explored a milder assumption: the availability of proxy variables to the latent confounding variables. Miao et al. (2018) demonstrated in the *Proxy Causal Learning* (PCL) framework that two proxy variables - a treatment proxy (possibly directly causally linked to the treatment) and an outcome proxy (possibly directly causally linked to the outcome) - are sufficient for recovering the underlying causal relation by utilizing an *outcome bridge function*, without the need to explicitly recover the confounder (unlike Kuroki and Pearl, 2014, who explicitly recover the confounder in the discrete-valued categorical setting). The corresponding causal graph is illustrated in Figure 1. An alternative line of research has proposed methods for causal effect estimation by leveraging a *treatment bridge function* for discrete treatment (Cui et al., 2024) and continuous treatment (Deaner, 2023; Wu et al., 2024), and our approach follows this setting. Specifically, Cui et al. (2024); Deaner (2023) introduced a bridge function $\varphi_0(Z, a)$ that satisfies $\mathbb{E}[\varphi_0(Z, a)|W, A = a] = 1/p(A = a|W)$, where Z denotes the treatment proxy, A denotes the treatment, and W denotes the outcome proxy, as illustrated in Figure 1. They showed that the dose-response can be identified through the expectation $\mathbb{E}[Y\varphi_0(Z, A)\mathbf{1}[A = a]]$. While the method of Cui et al. (2024) is limited to discrete treatments, Wu et al. (2024) address the case of continuous treatments by replacing the indicator function with a kernel function $K(A - a)$, yielding a dose-response estimator of the form $\mathbb{E}[Y\varphi_0(Z, a)K(A-a)]$; and by using a kernel density estimation or conditional normalizing flows to approximate $p(A = a|W)$. Kernel density estimates can converge slowly when the treatment and proxy variables are high-dimensional, however (see e.g. Wasserman, 2006).

In the present work, we propose a treatment proxy approach that eliminates the need for explicit density estimation. By leveraging a slightly different bridge function, we simplify the loss function by decomposing it into terms that depend on distinct distributions, akin to the approach of Kanamori et al. (2009), allowing us to express all quantities of interest in terms of inner products in *reproducing kernel Hilbert Spaces* (RKHS), removing the need for density ratio estimation. We further extend our approach to the conditional dose-response curve. In summary, our main contributions are as follows:

- We propose a novel family of kernel-based algorithms to estimate causal functions in the PCL setting, using a treatment bridge function;
- Our RKHS formulation allows us to provide closed-form expressions for causal effect estimation, including for continuous and high-dimensional treatments;
- We prove the consistency of proposed estimators;
- We demonstrate that our treatment proxy approach matches or outperforms existing PCL algorithms.

The paper is organized as follows: Section (2) reviews related work, Section (3) presents the problem definition and identification results, and Section (4) outlines our estimation algorithms. Consistency results are in Section (5), followed by experiments in Section (6).

2 RELATED WORK

Proxy causal learning (Miao et al., 2018; Deaner, 2023), building on the seminal work of Kuroki and Pearl (2014), tackles the problem of unmeasured confounding by using two types of proxy variables, namely *the treatment proxy* and *the outcome proxy*. Given these proxies, there are two approaches to obtain the causal effect. One approach is to estimate an *outcome bridge function*. This is a function of the outcome proxy, and we can obtain the causal effect by integrating the outcome bridge over the (observed) outcome proxy. Miao et al. (2018) show that the outcome bridge function is the solution of an inverse problem known as a Fredholm integral equation of the first kind (Kress, 2013). Although it is ill-posed in general, a number of methods have been proposed to solve it by limiting the functional space, including sieve bases (Deaner, 2023), RKHSs (Mastouri et al., 2021; Singh, 2023) and neural networks (Xu et al., 2021; Kompa et al., 2022; Kallus et al., 2021). The other approach, known as *alternative PCL*, considers a *treatment bridge function*. This is a function of the treatment proxy, which can be used as the adjustment weights in estimating causal effects, similar to the inverse propensity score (Rosenbaum and Rubin, 1983). Such a function can be obtained by solving another Fredholm integral equation (Deaner, 2023; Cui et al., 2024), but it is more challenging since it involves the conditional density function. Recently, Wu et al. (2024) proposed a “plug-in” approach, which explicitly performs density estimation in obtaining a treatment bridge. However, conditional density estimation is costly and suffers from slow convergence when the treatment is high-dimensional. Instead, we propose to bypass the need for density estimation using the conditional kernel mean embedding (Song et al., 2009; Grünewälder et al., 2012a; Park and Muandet, 2020; Klebanov et al., 2020; Li et al., 2022).

Proxy methods are also used in domain adaptation under distribution shifts (Alabdulmohsin et al., 2023; Tsai et al., 2024), where the underlying DAG resembles ours. Domain adaptation focuses on transferring models across domains with shifting unobserved confounders (e.g., patients in different hospitals). Tsai et al. (2024) use kernel-based outcome-bridge function similar to (Mastouri et al., 2021), whereas we employ treatment-bridge functions for causal effect estimations under unmeasured confounding. While both they and we use kernel methods, their objectives and estimands differ, highlighting distinct but complementary approaches.

3 PROBLEM SETTING AND IDENTIFICATION

3.1 Problem Setting for Treatment Effects

In this section, we establish the problem setting for learning causal effects, which are statements about the counterfactual outcomes that arise from hypothetical interventions. Consider the treatment $A \in \mathcal{A}$ and its associated outcome $Y \in \mathbb{R}$ that we observe. We assume the existence of an unobserved confounding variable $U \in \mathcal{U}$ that affects both A and Y. Our aim is to estimate the causal effects presented in the following definition.

Definition 3.1. *The treatment effects are:*

i-) Dose-response: $f_{ATE}(a) = \mathbb{E}[\mathbb{E}[Y|A = a, U]]$ quantifies the counterfactual mean outcome across the entire population given the intervention where everyone receives the treatment a. ATE signifies that its semiparametric analogue is the average treatment effect.

ii-) Conditional dose-response: $f_{ATT}(a, a') = \mathbb{E}[\mathbb{E}[Y|A = a, U]|A = a']$ quantifies the counterfactual mean outcome for individuals who actually received treatment $A = a'$ given the intervention where they receive treatment a. ATT signifies that its semiparametric analogue is the average treatment effect on the treated.

The primary challenge in estimating these functions is that the confounder U is not directly observable. To address this issue, we assume access to two proxy variables: Z, the treatment proxy, and W, the outcome proxy. We list the assumptions below that will be used throughout the development of our method. The following conditional independence assumption is implied by the graphical model in Figure (1).

Assumption 3.2. *We assume the following conditional independence statements: i-) $Y \perp Z|U, A$ (Conditional Independence for Y), ii-) $W \perp Z|U, A$ and $W \perp A|U$ (Conditional Independence for W).*

We also make the following completeness assumption.

Assumption 3.3 (Completeness)**.** *For any square integrable function $\ell : \mathcal{U} \to \mathbb{R}$, for all $a \in \mathcal{A}$: $\mathbb{E}[\ell(U)|W, A = a] = 0$ $p(W)-$almost everywhere (a.e.) if and only if $\ell(U) = 0$ $p(U)-$a.e.*

Both Assumptions (3.2) and (3.3) are used in the alternative PCL frameworks (Deaner, 2023; Cui et al., 2024; Wu et al., 2024). The completeness condition ensures that the proxy variable W has sufficient variability relative to the unobserved confounder U, thereby making it possible to identify the treatment effects.

3.2 Identification of the Structural Functions

In this section, we establish the identifiability of the structural functions presented in Definition (3.1). To obtain the ATE function given in Definition (3.1-i), we consider the *bridge function* $\varphi_0^{\mathrm{ATE}}(z, a)$, defined as a solution of the functional equation

$$\mathbb{E}[\varphi_0^{\mathrm{ATE}}(Z, a)|W, A = a] = \frac{p(W)p(a)}{p(W, a)}. \tag{1}$$

The densities $p(W)$, $p(A)$, and $p(W, A)$ denote the marginal distributions of W and A, and the joint distribution of (W, A), respectively. Similar to ATE, we consider a bridge function $\varphi_0^{\mathrm{ATT}}(z, a, a')$ to identify f_{ATT} that satisfies the functional equation

$$\mathbb{E}[\varphi_0^{\mathrm{ATT}}(Z, a, a')|W, A = a] = \frac{p(W, a')p(a)}{p(W, a)p(a')}. \tag{2}$$

In the following theorem, we establish the identifiability of the structural functions f_{ATE} and f_{ATT}.

Theorem 3.4. *Let Assumptions (3.2) and (3.3) hold. Furthermore, suppose that there exist square integrable functions φ_0^{ATE} and φ_0^{ATT} that satisfy Equations (1) and (2), respectively. Then,*

1. *The dose-response curve is given by $f_{ATE}(a) = \mathbb{E}[Y\varphi_0^{ATE}(Z, a)|A = a]$.*
2. *The conditional dose-response curve is given by $f_{ATT}(a, a') = \mathbb{E}[Y\varphi_0^{ATT}(Z, a, a')|A = a]$.*

Theorem (3.4) is proved in the supplementary material S.M. (Section 9). The extension to settings with additional observable confounders is in S.M. (Sec. 15).

Remark 3.5. *Our ATE identification result differs from previous works (Wu et al., 2024; Cui et al., 2024). Specifically, (Wu et al., 2024) extends (Cui et al., 2024) by replacing the indicator function with a kernel for continuous treatments, identifying ATE via expectations over $p(Y, Z, A)$. By contrast, our approach uses conditional expectations over $p(Y, Z|A =$*

a), enabled by a modified bridge function definition with an additional $p(A = a)$ term. This distinction highlights the novelty of our approach to ATE identification. Furthermore, unlike (Wu et al., 2024), our approach does not require density ratio estimation for bridge function estimation, as outlined in Section (4).

Remark 3.6. *For the simpler case where $\{\mathcal{A}, \mathcal{Y}, \mathcal{W}, \mathcal{Z}, \mathcal{U}\}$ are discrete, we present the identification result in Theorem (14.1). Specifically, in this setting, the dose-response can be identified through matrix-vector multiplication of probability matrices, which can be estimated from the variables (A, Y, W, Z), allowing for the estimation of the dose-response curve.*

In order to ensure that solutions exist for both ATE and ATT bridge functions, we require the following completeness assumption:

Assumption 3.7. *For any square integrable function $\ell : \mathcal{U} \to \mathbb{R}$, for all $a \in \mathcal{A}$: $\mathbb{E}[\ell(U)|Z, A = a] = 0$ $p(Z)-$a.e. if and only if $\ell(U) = 0 \quad p(U)$-a.e.*

The assumption above, along with mild regularity and integrability conditions, suffices for the existence of solutions to Equations (1) and (2). Further discussions about the existence are provided in S.M. (Sec. 10).

4 METHODS

With the identification results of ATE and ATT functions in hand, we are prepared to develop algorithms to estimate these structural functions. To achieve this, we must solve Equations (1) and (2). We assume that the bridge functions reside within RKHSs. We then use a two-stage regression approach to solve for the bridge functions. Ultimately, we derive closed-form solutions to estimate the structural functions.

4.1 Reproducing Kernel Hilbert Space

Consider any space $\mathcal{F} \in \{\mathcal{A}, \mathcal{W}, \mathcal{Z}\}$. We denote the positive semi-definite kernel on $\mathcal{F}$ as $k_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$. The corresponding canonical feature map is $\phi_{\mathcal{F}}(f)$, where $\phi_{\mathcal{F}}(f) = k_{\mathcal{F}}(f, \cdot) \in \mathcal{H}_{\mathcal{F}}$, and $\mathcal{H}_{\mathcal{F}}$ refers to the RKHS of real-valued functions defined on $\mathcal{F}$. The inner product and the norm of the RKHS are denoted by $\langle ., . \rangle_{\mathcal{H}_{\mathcal{F}}}$ and $\|.\|_{\mathcal{H}_{\mathcal{F}}}$, respectively. When it is clear from the context, we drop the subscript $\mathcal{H}_{\mathcal{F}}$ from the inner product notation. For notational convenience, we use $\mathcal{H}_{\mathcal{F}\mathcal{G}}$ to denote the tensor product space $\mathcal{H}_{\mathcal{F}} \otimes \mathcal{H}_{\mathcal{G}}$ that is isometrically isomorphic to $S_2(\mathcal{H}_{\mathcal{G}}, \mathcal{H}_{\mathcal{F}})$ the Hilbert space of Hilbert-Schmidt operators from $\mathcal{H}_{\mathcal{G}}$ to $\mathcal{H}_{\mathcal{F}}$ (Aubin, 2011). Correspondingly, $\phi_{\mathcal{F}\mathcal{G}}(f, g)$ denotes the tensor product feature map $\phi_{\mathcal{F}}(f) \otimes \phi_{\mathcal{G}}(g)$. Given any distribution $p(F)$ on $\mathcal{F}$ and a kernel $k_{\mathcal{F}}$ for which $\mathbb{E}[k_{\mathcal{F}}(F, F)] < \infty$, $\mu_F = \int \phi_{\mathcal{F}}(f) p(f) df \in \mathcal{H}_{\mathcal{F}}$ is known as the kernel mean embedding of p (Smola et al., 2007; Gretton, 2013). Similarly, for a conditional distribution $p(F|g)$ for each $g \in \mathcal{G}$, the operator $\mu_{F|G}(g) = \int \phi_{\mathcal{F}}(f) p(f|g) df \in \mathcal{H}_{\mathcal{F}}$ is called the conditional mean embedding (CME) of $p(F|g)$ (Song et al., 2009; Grünewälder et al., 2012a; Park and Muandet, 2020; Klebanov et al., 2020; Li et al., 2022).

4.2 Algorithms to Estimate Causal Functions

4.2.1 Dose-Response Estimation

To estimate the dose-response curve, we first approximate the bridge function φ_0^{ATE}, which is defined as the solution of Equation (1). Let $r(W, A)$ denote $p(W)p(A)/p(W, A)$. Our objective is to find the optimal solution to the least-squares loss $\mathbb{E}[(r(W, A) - \mathbb{E}[\varphi(Z, A)|W, A])^2]$. Minimizing this loss is challenging for two reasons: (i) it requires knowledge of the density ratios; (ii) it involves a conditional expectation. For the first challenge, one could perform density ratio estimation, but this is particularly difficult in high-dimensional settings. We avoid the need for density ratio estimation by simplifying the least-squares objective as follows:

$$
\begin{aligned}
&\mathbb{E}\big[\big(r(W, A) - \mathbb{E}[\varphi(Z, A)|W, A]\big)^2\big] \\
&= \mathbb{E}\big[\mathbb{E}[\varphi(Z, A)|W, A]^2\big] \\
&- \int \frac{2p(w)p(a)}{p(w, a)} \mathbb{E}[\varphi(Z, a)|w, a] p(w, a) dw da + \text{const.} \\
&= \mathbb{E}\big[\mathbb{E}[\varphi(Z, A)|W, A]^2\big] \\
&- \ 2\mathbb{E}_W \mathbb{E}_A \big[\mathbb{E}[\varphi(Z, A)|W, A]\big] + \text{const.} \qquad (3)
\end{aligned}
$$

Here, $\mathbb{E}_W \mathbb{E}_A[.]$ denotes decoupled expectations with respect to $p(w)$ and $p(a)$, i.e., for any function function ℓ, $\mathbb{E}_W \mathbb{E}_A[\ell(W, A)] = \int \ell(w, a) p(w) p(a) dw da$, and $'$const.$'$ represents terms independent of φ. The above simplification allows us to avoid the need to estimate a density ratio, since minimizing Equation (3) does not require knowing $r(W, A)$. To overcome the second challenge, we assume that φ_0^{ATE} resides in the RKHS $\mathcal{H}_{\mathcal{Z}\mathcal{A}}$. Observe that for any $\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}}$,

$$
\begin{aligned}
&\mathbb{E}[\varphi(Z, a)|W = w, A = a] \\
&= \mathbb{E}[\langle \varphi, \phi_{\mathcal{Z}}(Z) \otimes \phi_{\mathcal{A}}(a) \rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} | W = w, A = a] \\
&= \langle \varphi, \mathbb{E}[\phi_{\mathcal{Z}}(Z)|W = w, A = a] \otimes \phi_{\mathcal{A}}(a) \rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \\
&= \langle \varphi, \mu_{Z|W,A}(w, a) \otimes \phi_{\mathcal{A}}(a) \rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}},
\end{aligned}
$$

where $\mu_{Z|W,A}(w, a)$ denotes the CME $\mathbb{E}[\phi_{\mathcal{Z}}(Z)|W = w, A = a]$. With this further simplification, we can write Equation (3) as

$$
\begin{aligned}
&\mathbb{E}\big[\langle \varphi, \mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \rangle^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\big] \\
&- 2\mathbb{E}_W \mathbb{E}_A \big[\langle \varphi, \mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\big] + \text{const.}
\end{aligned}
$$

As a result, our approach to estimate the bridge function will be a *two-stage procedure*: (i) The first stage estimates the CME $\mu_{Z|W,A}(W, A)$; (ii) the second stage minimizes the modified regression loss using the approximated CME. Since we identify the dose-response through a conditional expectation, our approach requires learning an additional conditional mean embedding, as outlined below, which we consider as a third-stage regression.

First Stage Regression: Under the regularity condition that $\mathbb{E}[\ell(Z)|W = \cdot, A = \cdot] \in \mathcal{H}_{\mathcal{WA}}$ for all $\ell \in \mathcal{H}_{\mathcal{Z}}$, there exist an operator $C_{Z|W,A} \in S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})$ such that $\mu_{Z|W,A}(w, a) = C_{Z|W,A}\phi_{\mathcal{WA}}(w, a)$ (Song et al., 2009; Li et al., 2024). $C_{Z|W,A}$ is called the CME operator and can be estimated by vector-valued regression (Grünewälder et al., 2012a; Mollenhauer and Koltai, 2020; Li et al., 2022, 2024). The estimation of the CME operator is referred to as the *first-stage regression.* Given first-stage samples $\{w_i, a_i, z_i\}_{i=1}^n$, $C_{Z|W,A}$ is learned by minimizing the regularized vector-valued least-squares cost

$$\hat{\mathcal{L}}^c(C) = \frac{1}{n}\sum_{i=1}^{n} \|\phi_{\mathcal{Z}}(z_i) - C\phi_{\mathcal{WA}}(w_i, a_i)\|^2_{\mathcal{H}_{\mathcal{Z}}} + \lambda_1 \|C\|^2_{S_2}$$

i.e., $\hat{C}_{Z|W,A} = \arg\min_{C \in S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})} \hat{\mathcal{L}}^c(C)$. The solution to this problem is given by $\hat{C}_{Z|W,A} = \hat{C}_{Z,(W,A)}(\hat{C}_{W,A} + \lambda_1 I)^{-1}$ (Grünewälder et al., 2012b), where $\hat{C}_{Z,(W,A)} = \frac{1}{n}\sum_{i=1}^n \phi_{\mathcal{Z}}(z_i) \otimes \phi_{\mathcal{WA}}(w_i, a_i)$ and $\hat{C}_{W,A} = \frac{1}{n}\sum_{i=1}^n \phi_{\mathcal{WA}}(w_i, a_i) \otimes \phi_{\mathcal{WA}}(w_i, a_i)$.

Second Stage Regression: Given the CME estimation $\hat{\mu}_{Z|W,A}(w, a) = \hat{C}_{Z|W,A}\phi_{\mathcal{WA}}(w, a)$ from the first-stage, we aim to minimize the modified least-squares loss. We minimize the empirical counterpart of this loss with Tikhonov regularization, using the CME estimate $\hat{\mu}_{Z|W,A}$ and the second-stage data $\{\tilde{z}_i, \tilde{w}_i, \tilde{a}_i\}_{i=1}^m$. The empirical loss $\hat{\mathcal{L}}_m^{2SR}(\varphi)$ is expressed as

$$\hat{\mathcal{L}}_m^{2SR}(\varphi) = \frac{1}{m}\sum_{i=1}^{m} \langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\rangle^2_{\mathcal{H}_{\mathcal{ZA}}} - \frac{2}{m(m-1)} \sum_{\substack{i,j=1 \\ j \neq i}}^{m} \Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle_{\mathcal{H}_{\mathcal{ZA}}} + \lambda_2 \|\varphi\|^2_{\mathcal{H}_{\mathcal{ZA}}}$$

We denote its minimizer on $\mathcal{H}_{\mathcal{ZA}}$ as $\hat{\varphi}_{\lambda_2,m}$. This is referred to as the *second-stage regression.*

Third Stage Regression: We can then estimate the ATE function as $f_{\text{ATE}}(a) \approx \mathbb{E}[Y \hat{\varphi}_{\lambda_2,m}(Z, a)|A = a]$ in closed-form via kernel matrices. We note that:

$$\begin{aligned}
&\mathbb{E}[Y \hat{\varphi}_{\lambda_2,m}(Z, a)|A = a] \\
&\quad = \mathbb{E}[Y \langle \hat{\varphi}_{\lambda_2,m}, \phi_{\mathcal{Z}}(Z) \otimes \phi_{\mathcal{A}}(a)\rangle|A = a] \\
&\quad = \langle \hat{\varphi}_{\lambda_2,m}, \mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A = a] \otimes \phi_{\mathcal{A}}(a)\rangle.
\end{aligned}$$

The evaluation of the above inner product requires the estimation of $\mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A = a]$. We obtain this via kernel ridge regression using data from either first-stage, second-stage, or a combination of both. This can be considered as a *third-stage regression.* Similar to the first-stage regression, we estimate the conditional mean operator $C_{YZ|A}$ such that $C_{YZ|A}\phi_{\mathcal{A}}(a) = \mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A = a]$. For simplicity in our algorithm derivations, we use first-stage data to estimate this conditional mean. Replacing this conditional mean with its estimate, we can estimate the dose-response with the inner product $\hat{f}_{\text{ATE}}(a) = \langle \hat{\varphi}_{\lambda_2,m}, \hat{\mathbb{E}}[Y\phi_{\mathcal{Z}}(Z)|A = a] \otimes \phi_{\mathcal{A}}(a)\rangle$. The algorithm below provides the closed-form solution for ATE estimation as a result of the three stages, and its derivation can be found in S.M. (Sec. 11.1).

Algorithm 4.1. *Let the first and second-stage data be denoted by* $\{z_i, w_i, a_i\}_{i=1}^n$ *and* $\{\tilde{w}_i, \tilde{a}_i\}_{i=1}^m$*, respectively, and* $(\lambda_1, \lambda_2, \lambda_3)$ *be the regularization parameters. For* $F \in \{A, W, Z\}$ *with domain* $\mathcal{F}$*, the first-stage kernel matrices are denoted as* $\boldsymbol{K}_{FF} = [k_{\mathcal{F}}(f_i, f_j)]_{ij} \in \mathbb{R}^{n\times n}$, $\boldsymbol{K}_{Ff} = [k_{\mathcal{F}}(f_i, f)]_i \in \mathbb{R}^n$*, where* $\{f_i\}_{i=1}^n$ *denotes the first-stage data samples. For the second-stage variables* $\tilde{F} \in \{\tilde{A}, \tilde{W}\}$*, the kernel matrices are denoted as:* $\boldsymbol{K}_{\tilde{F}\tilde{F}} = [k_{\mathcal{F}}(\tilde{f}_i, \tilde{f}_j)]_{ij} \in \mathbb{R}^{m\times m}$, $\boldsymbol{K}_{F\tilde{F}} = [k_{\mathcal{F}}(f_i, \tilde{f}_j)]_{ij} \in \mathbb{R}^{n\times m}$, $\boldsymbol{K}_{F\tilde{f}} = [k_{\mathcal{F}}(f_i, \tilde{f})]_i \in \mathbb{R}^n$*, and* $\boldsymbol{K}_{\tilde{F}f} = [k_{\mathcal{F}}(\tilde{f}_j, f)]_j \in \mathbb{R}^m$*. Define the following matrices: i-)* $\boldsymbol{B} = (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1\boldsymbol{I})^{-1}(\boldsymbol{K}_{W\tilde{W}} \odot \boldsymbol{K}_{A\tilde{A}}) \in \mathbb{R}^{n\times m}$*, ii-)* $\bar{\boldsymbol{B}} \in \mathbb{R}^{n\times m}$ *is the matrix, where j-th column is given by* $\bar{\boldsymbol{B}}_{:,j} = \frac{1}{m}\sum_{\substack{l=1 \\ l\neq j}}^{m}(\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1\boldsymbol{I})^{-1}(\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j})$*, where* $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ *is the identity matrix. Furthermore, let* $\{\alpha_i\}_{i=1}^{m+1}$ *be the minimizer of the cost function* $\hat{\mathcal{L}}_m^{2SR}(\boldsymbol{\alpha}) = \frac{1}{m}\boldsymbol{\alpha}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\boldsymbol{M} + \lambda_2\boldsymbol{\alpha}^T\boldsymbol{N}\boldsymbol{\alpha}$ *where*

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{B}^T\boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\boldsymbol{1}}{m})^T\big[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T \end{bmatrix}^T \in \mathbb{R}^{m\times(m+1)},$$

$$\boldsymbol{M} = \begin{bmatrix} [\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]\frac{\boldsymbol{1}}{m} \\ (\frac{\boldsymbol{1}}{m})^T\Big[\bar{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big]\frac{\boldsymbol{1}}{m} \end{bmatrix} \in \mathbb{R}^{(m+1)},$$

$\boldsymbol{N} = \begin{bmatrix}\boldsymbol{L} & \boldsymbol{M}\end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$, $\boldsymbol{1} \in \mathbb{R}^m$ *is vector of ones, and* $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_m & \alpha_{m+1}\end{bmatrix}^T \in \mathbb{R}^{m+1}$*. Then the dose-response estimation can be written in closed-form as* $\hat{f}_{ATE}(a) = \boldsymbol{\alpha}^T\boldsymbol{E}$*, where*

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{B}^T\boldsymbol{D}\boldsymbol{K}_{Aa} \odot \boldsymbol{K}_{\tilde{A}a} \\ (\bar{\boldsymbol{B}}^T\boldsymbol{D}\boldsymbol{K}_{Aa} \odot \boldsymbol{K}_{\tilde{A}a})\frac{\boldsymbol{1}}{m} \end{bmatrix} \in \mathbb{R}^{m+1},$$

$\boldsymbol{D} = \boldsymbol{K}_{ZZ}\,diag(\boldsymbol{Y})[\boldsymbol{K}_{AA} + n\lambda_3\boldsymbol{I}]^{-1} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{Y} = \begin{bmatrix}y_1 & y_2 & \dots & y_n\end{bmatrix}^T \in \mathbb{R}^n$.

4.2.2 Conditional Dose-Response Estimation

As with the dose-response, we aim to minimize the least-squares loss $\mathbb{E}_{W,A}[(r(W,A,a') - \mathbb{E}[\varphi(Z,A,a')|W,A])^2]$ for conditional dose-response. Our method allows for similar simplifications in the second-stage regression as before, thereby bypassing explicit density estimation in minimizing the loss function $\mathcal{L}^{\text{2SR}}$. Let $r(W,A,a') = \frac{p(W,a')p(A)}{p(W,A)p(a')}$ and suppose that the bridge function φ lies in the RKHS $\mathcal{H}_{\mathcal{ZAA}}$ (which denotes $\mathcal{H}_\mathcal{Z} \otimes \mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{A}$). Specifically, the loss function can be simplified as (see S.M., Sec. 11.2):

$$\begin{aligned}&\mathbb{E}[(r(W,A,a') - \mathbb{E}[\varphi(Z,A,a')|W,A])^2]\\ &= \mathbb{E}[\langle \varphi, \mu_{Z|W,A}(W,A) \otimes \phi_\mathcal{A}(A) \otimes \phi_\mathcal{A}(a')\rangle^2]\\ &- 2\mathbb{E}_A[\langle \varphi, C_{Z|W,A}(\mathbb{E}_{W|A=a'}[\phi_\mathcal{W}(W)] \otimes \phi_\mathcal{A}(A))\\ &\otimes \phi_\mathcal{A}(A) \otimes \phi_\mathcal{A}(a')\rangle] + \text{const.}\end{aligned}$$

The notation $\mathbb{E}_{W|A=a'}$ denotes the expectation with respect to the conditional distribution $p(W|A = a')$, i.e., for any function ℓ, $\mathbb{E}_{W|A=a'}[\ell(W)] = \int \ell(w)p(w|a')dw$. We need to estimate the CME $\mathbb{E}[\phi_\mathcal{W}(W)|A = a']$, which can be obtained using kernel ridge regression on second-stage data, and expressed in closed form as $\hat{\mathbb{E}}[\phi_\mathcal{W}(W)|A = a'] = \sum_{i=1}^m \theta_i \phi_\mathcal{W}(\tilde{w}_i) = \Phi_\mathcal{W}\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\boldsymbol{K}_{\tilde{A}\tilde{A}} + m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{A}a'}$, $\Phi_\mathcal{W} = \begin{bmatrix}\phi_\mathcal{W}(\tilde{w}_1) & \dots & \phi_\mathcal{W}(\tilde{w}_m)\end{bmatrix}$, and ζ is the regularization parameter for this CME estimate. Hence, we can write the sample-based loss, $\hat{\mathcal{L}}_m^{\text{2SR}}(\varphi)$, of the second-stage regression with Tikhonov regularization,

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m \langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_\mathcal{A}(\tilde{a}_i) \otimes \phi_\mathcal{A}(a')\rangle^2\\ &- \frac{2}{m}\sum_{\substack{i,j=1\\ i\neq j}}^m \langle \varphi, \theta_i \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_\mathcal{A}(\tilde{a}_j) \otimes \phi_\mathcal{A}(a')\rangle\\ &+ \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{ZAA}}}.\end{aligned}$$

Using the estimate $\hat{\varphi}_{\lambda_2,m}$ for φ_0^{ATT} from the second stage regression, the conditional dose-response curve is given by the inner product $\hat{f}_{\text{ATT}}(a,a') = \langle \hat{\varphi}_{\lambda_2,m}, \hat{\mathbb{E}}[Y\phi_\mathcal{Z}(Z)|A = a] \otimes \phi_\mathcal{A}(a) \otimes \phi_\mathcal{A}(a')\rangle$, where we again need the conditional mean $\hat{\mathbb{E}}[Y\phi_\mathcal{Z}(Z)|A = a]$ as for the dose-response algorithm. The algorithm below provides a closed-form solution for the conditional dose-response estimate, with proof in S.M. (Sec. 11.2).

Algorithm 4.2. *Denote the first- and second-stage data as $\{z_i, w_i, a_i\}_{i=1}^n$ and $\{\tilde{w}_i, \tilde{a}_i\}_{i=1}^m$, respectively, and let $(\lambda_1, \lambda_2, \lambda_3, \zeta)$ be the regularization parameters. Define the kernel matrices, the matrix $\boldsymbol{D} \in \mathbb{R}^{n\times n}$, and the matrix $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ as in Algorithm (4.1). Furthermore, define $\tilde{\boldsymbol{B}}$, where j-th column is given by $\tilde{\boldsymbol{B}}_{:,j} = \sum_{\substack{l=1\\ l\neq j}}^m(\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1\boldsymbol{I})^{-1}(\theta_l\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j})$ with $\theta_i = [(\boldsymbol{K}_{\tilde{A}\tilde{A}} + m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{A}a'}]_i$. For a given a', let $\{\alpha_i\}_{i=1}^{m+1}$ be the minimizer of $\hat{\mathcal{L}}_m^{2SR}(\boldsymbol{\alpha}) = k_\mathcal{A}(a',a')\big(\frac{k_\mathcal{A}(a',a')}{m}\boldsymbol{\alpha}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T\boldsymbol{M} + \lambda_2\boldsymbol{\alpha}^T\boldsymbol{N}\boldsymbol{\alpha}\big)$ where*

$$\boldsymbol{L} = \begin{bmatrix}\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\\ (\frac{\boldsymbol{1}}{m})^T\big[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T\end{bmatrix}^T \in \mathbb{R}^{m\times(m+1)},$$

$$\boldsymbol{M} = \begin{bmatrix}[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]\frac{\boldsymbol{1}}{m}\\ (\frac{\boldsymbol{1}}{m})^T\Big[\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big]\frac{\boldsymbol{1}}{m}\end{bmatrix} \in \mathbb{R}^{(m+1)},$$

and $\boldsymbol{N} = \begin{bmatrix}\boldsymbol{L} & \boldsymbol{M}\end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)}$. Then, the conditional dose-response estimate can be written in closed-form as $\hat{f}_{ATT}(a) = k_\mathcal{A}(a',a')\boldsymbol{\alpha}^T\boldsymbol{E}$, where

$$\boldsymbol{E} = \begin{bmatrix}\boldsymbol{B}^T\boldsymbol{D}\boldsymbol{K}_{Aa} \odot \boldsymbol{K}_{\tilde{A}a}\\ (\tilde{\boldsymbol{B}}^T\boldsymbol{D}\boldsymbol{K}_{Aa} \odot \boldsymbol{K}_{\tilde{A}a})\frac{\boldsymbol{1}}{m}\end{bmatrix} \in \mathbb{R}^{m+1}.$$

5 CONSISTENCY

We present non-asymptotic uniform consistency guarantees for the dose-response curve; similar guarantees for the conditional dose-response curve are provided in S.M. (Sec. 12). We recall that for the third-stage regression, we can re-use data from the first and second stages, and we denote by t the number of samples used for that stage; thus n, m, t are the number of samples used in stages 1, 2, and 3, respectively. Likewise $\lambda_1, \lambda_2, \lambda_3$ are the regularization parameters for their respective stages. We assume that each regression stage is well specified, as follows:

Assumption 5.1. *(1) There exists $C_{Z|W,A} \in S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_\mathcal{Z})$ such that $\mu_{Z|W,A}(W,A) = C_{Z|W,A}\phi_{\mathcal{WA}}(W,A)$; (2) There exists a solution $\varphi_0 \in \mathcal{H}_{\mathcal{ZA}}$ of Equation (1); (3) There exists $C_{YZ|A} \in S_2(\mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{Z})$ such that $\mathbb{E}[Y\phi_\mathcal{Z}(Z)|A] = C_{YZ|A}\phi_\mathcal{A}(A)$.*

We impose the following additional conditions.

Assumption 5.2. *For $\mathcal{F} \in \{\mathcal{A}, \mathcal{W}, \mathcal{Z}\}$, we assume*

i. $\mathcal{F}$ is a Polish space;

ii. $k_\mathcal{F}(f,.)$, is continuous for almost every $f \in \mathcal{F}$ and is also bounded by κ for almost every $f \in \mathcal{F}$, i.e., $\sup_{f\in\mathcal{F}}\|k_\mathcal{F}(f,.)\|_{\mathcal{H}_\mathcal{F}} \leq \kappa$;

iii. There exists $R, \sigma > 0$ such that $\forall\ q \geq 2$, P_A-almost surely, $\mathbb{E}[(Y - \mathbb{E}[Y \mid A])^q \mid A] \leq \frac{1}{2}q!\sigma^2R^{q-2}$.

Assumption 5.3. *Let $\bar{\varphi}_0$ be the minimum RKHS norm bridge function solution from Definition (12.2), and let Σ_1, Σ_2, and Σ_3 be covariance operators associated with first, second, and third-stage regressions, respectively, as defined in Definition (12.4). We assume that the following conditions hold:*

i. There exists a constant $B_1 < \infty$ such that for a given $\beta_1 \in (1,3]$, $\|C_{Z|W,A}\Sigma_1^{-\frac{\beta_1-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{WA}},\mathcal{H}_\mathcal{Z})} \leq B_1$

ii. There exists a constant $B_2 < \infty$ such that for a given $\beta_2 \in (1,3]$, $\|\Sigma_2^{-\frac{\beta_2-1}{2}} \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \leq B_2$.

iii. There exists a constant $B_3 < \infty$ such that for a given $\beta_3 \in (1,3]$, $\|C_{YZ|A}\Sigma_3^{-\frac{\beta_3-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})} \leq B_3$.

Note that Assumption (3.7) implies that there exists a solution of Equation (1), Assumption (5.1-2) further requires that at least one solution lies in the RKHS. The Bernstein condition in Assumption (5.2-iii) regulates observation noise, while the source condition in Assumption (5.3) links regression smoothness to covariance operators (Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020). Theorem (5.4) also assumes an eigenvalue decay condition in Assumption (12.9) to characterize RKHSs' effective dimension.

Theorem 5.4. *Let Assumptions (3.2), (3.3), (3.7), (5.1), (5.2), (5.3) and (12.9) hold with parameters $\beta_1, \beta_2, \beta_3 \in (1,3]$ and $p_1, p_2, p_3 \in (0,1]$. Set $\lambda_1 = n^{-\frac{1}{\beta_1+p_1}}$ and $\lambda_3 = t^{-\frac{1}{\beta_3+p_3}}$. Fix $\iota > 0$ and $n = m^{\iota\frac{\beta_1+p_1}{\beta_1-1}}$.*

i. If $\iota \leq \frac{\beta_2+1}{\beta_2+p_2}$, let $\lambda_2 = m^{-\frac{\iota}{\beta_2+1}}$, then
$$\|\hat{f}_{ATE} - f_{ATE}\|_\infty = O_p\left(t^{-\frac{1}{2}\frac{\beta_3-1}{\beta_3+p_3}} + m^{-\frac{\iota}{2}\frac{\beta_2-1}{\beta_2+1}}\right)$$

ii. If $\iota \geq \frac{\beta_2+1}{\beta_2+p_2}$, let $\lambda_2 = m^{-\frac{1}{\beta_2+p_2}}$, then
$$\|\hat{f}_{ATE} - f_{ATE}\|_\infty = O_p\left(t^{-\frac{1}{2}\frac{\beta_3-1}{\beta_3+p_3}} + m^{-\frac{1}{2}\frac{\beta_2-1}{\beta_2+p_2}}\right)$$

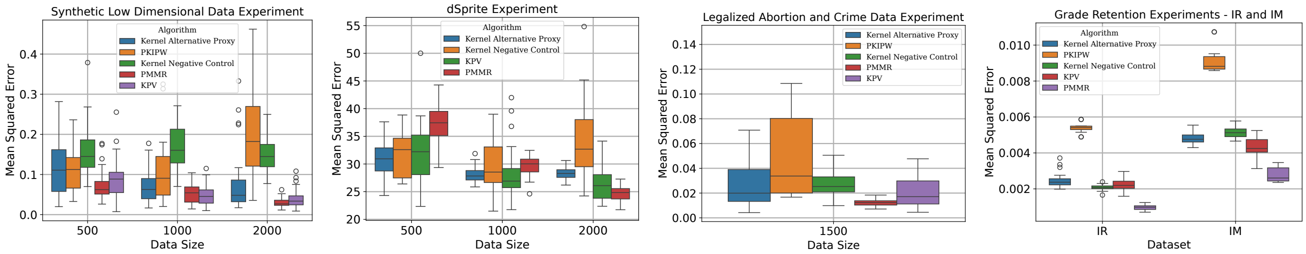
The proof and details on the assumptions are given in S.M. (Sec. 12). Parameters $\{p_i\}_{i=1}^3$ control the effective dimension of the RKHSs used in the three stages (Caponnetto and De Vito, 2007). A value $p_i \to 0$ corresponds to a finite dimensional RKHS, while larger p_i means slower decay of eigenvalues of the covariance operator and hence a larger effective dimension. Parameters $\{\beta_i\}_{i=1}^3$ control the smoothness of $C_{Z|W,A}$, φ_0 and $C_{YZ|A}$ respectively. Larger β_i corresponds to a smoother operator or function. ι controls the ratio between stage 1 and stage 2 samples to achieve a fast rate in the setting (ii). Indeed, at $\iota = (\beta_2+1)/(\beta_2+p_2)$, the convergence rate ($m^{-\frac{1}{2}\frac{\beta_2-1}{\beta_2+p_2}}$) is minimax optimal in m while requiring the fewest observations from stage 1 (Caponnetto and De Vito, 2007).

Comparison to outcome bridge function. Convergence guarantees for ATE when the outcome bridge technique is used have rate $t^{-1/2}$ instead of our rate $t^{-\frac{1}{2}\frac{\beta_3-1}{\beta_3+p_3}}$ (Mastouri et al. (Proposition 1 2021), Singh (Theorem 4 2023)). This is a consequence of the fact that our treatment bridge algorithm requires a third regression for stage 3, while the outcome bridge approach only requires an averaging for stage 3. While this poses a disadvantage in principle, it may be less important in practice than the ease of estimation of the treatment bridge vs the outcome bridge, in the same way that IPW and direct estimates may each be advantageous in different regimes (Bang and Robins, 2005). This can be observed in our experiments. Moreover, in Mastouri et al. (2021); Singh (2023), the rates stop improving at smoothness $\beta_i = 2$ while our rates improve up to $\beta_i = 3$. This improvement is obtained from a tighter control of the approximation error in RKHS norm, as observed by Meunier et al. (2023, Remark 7). This improvement can also be applied to previous works with an outcome bridge. Last, we emphasize that for consistency in the well-specified case, as treated by Mastouri et al. (2021); Singh (2023) and in our work, the kernels are not required to be characteristic (contrary to assumptions made in the earlier works), nor is Y required to be bounded (Y need only be sub-exponential).

Saturation effect. Benefits from high smoothness beyond the saturation point at $\beta_i = 3$ can be obtained by using alternative spectral regularization techniques (Engl et al., 1996). Results were recently obtained by Meunier et al. (2025) for conditional mean embedding learning. The application to the proxy setting is an interesting topic of future study.

6 NUMERICAL EXPERIMENTS

In this section, we assess the empirical performance of our proposed framework for estimating causal structural functions using both synthetic data and real-world tasks. We compare our method to other PCL frameworks, including Proximal Kernel Inverse Probability Weighted (PKIPW) (Wu et al., 2024), Kernel Negative Control (Singh, 2023), Proxy Maximum Moment Restriction (PMMR) (Mastouri et al., 2021) and Kernel Proxy Variable (KPV) (Mastouri et al., 2021). Additionally, for an ATT experiment, we compare our method to the Kernel-ATT algorithm proposed in (Singh et al., 2023), which assumes access to the confounding variable U. For each experiment (except those involving PKIPW), we employed a Gaussian kernel $k_{\mathcal{F}}(f_i, f_j) = \exp(-\|f_i - f_j\|_2^2/(2l^2))$ unless otherwise stated, where l is the length-scale of the kernel. The Gaussian kernel's length scale was selected using the median interpoint distance heuristic, if not specified otherwise. In the PKIPW experiments, we used the Epanechnikov kernel, consistent with the original implementation of Wu et al. (2024). To select the regularization parameters λ_1 and λ_3 (and ζ for the ATT) in our proposed methods, we employed leave-one-out cross-validation (LOOCV), which has a closed-form expression in the case of kernel ridge regression. For tuning the second-stage regularization parameter λ_2, we used the first-stage data as a held-out set to measure the validation loss, with added complexity regularization to avoid overfitting (Meanti et al., 2022).



(a) (b) (c) (d)

Figure 2: Dose-response curve estimation across various datasets and algorithms: *Kernel Alternative Proxy* (Ours), PKIPW (Wu et al., 2024), Kernel Negative Control (Singh, 2023), KPV (Mastouri et al., 2021), and PMMR (Mastouri et al., 2021). (a) Synthetic low-dimensional setting, (b) dSprite dataset, (c) legalized abortion and crime dataset, and (d) grade retention and cognitive outcome datasets.

Additional details, including ablation studies, hyperparameter selection procedures, and a GitHub link to our implementation can be found in S.M. (Sec. 13).

6.1 Dose-Response Experiments

We assess the performance of our proposed ATE algorithm on four datasets that are described below.

Low-Dimensional Setting: We use the data generation process outlined by (Wu et al., 2024), which incorporates a non-linear relationship between treatment and outcome:

$$
\begin{aligned}
&U_1 \sim \mathcal{U}[-1,2],\;\; U_2 \sim \mathcal{U}[0,1] - \mathbf{1}[0 \leq U_1 \leq 1], \\
&W = [U_2 + \mathcal{U}[-1,1], U_1 + \mathcal{N}(0,1)] \\
&Z = [U_2 + \mathcal{N}(0,1), U_1 + \mathcal{U}[-1,1]], A := U_1 + \mathcal{N}(0,1) \\
&Y := 3\cos(2(0.3U_2 + 0.3U_1 + 0.2) + 1.5A) + \mathcal{N}(0,1),
\end{aligned}
$$

where $\mathcal{U}[a,b]$ denotes the uniform distribution over the interval $[a,b]$, and $\mathcal{N}(\mu,\sigma^2)$ denotes Gaussian distribution with mean μ and variance σ^2. We use training sets of size 500, 1000, and 2000 in our experiments. Figure (2a) illustrates the mean squared error results, averaged over 30 realizations, comparing our method to other approaches. Our proposed method outperforms PKIPW and Kernel Negative Control, particularly as the size of the training data increases.

dSprite: The *Disentanglement testing Sprite dataset (dSprite)* contains images of size 64 × 64, described by latent parameters: *scale*, *rotation*, *posX*, and *posY* (Matthey et al., 2017). This dataset has been used to test the disentenglament properties of unsupervised models (Higgins et al., 2017). Xu and Gretton (2023) introduced benchmark dataset for PCL setting, where they treat the *dSprite* images as the high-dimensional treatments. Specifically, they consider the flattened image that is corrupted by Gaussian noise as the treatment, and the structural function of interest is defined as $f_{\text{ATE}}(A) = (\|BA\|_2^2 - 3000)/500$, where $A \in \mathbb{R}^{4096}$ and $B \in \mathbb{R}^{10\times 4096}$, with entries of B given by $B_{ij} = |32 - j|/32$. The outcome is defined by the expression $Y = 12(posY - 0.5)^2 f_{\text{ATE}}(A) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.5)$. The treatment inducing proxy $Z \in \mathbb{R}^3$ includes three variables: *scale*, *rotation*, *posX*. The outcome-inducing proxy is another *dSprite* image that shares the same *posY* variable as the treatment while the remaining variables are fixed as *scale* = 0.8. *rotation* = 0, and *posX* = 0.5. We use training sets of size 500, 1000, and 2000 in our experiments. Figure (2b) shows the mean squared error results averaged over 30 realizations, comparing our method to other approaches. As the code of (Wu et al., 2024) does not support the high dimensional treatment, we do not report results for PKIPW in this experiment. Our method outperforms Kernel Negative Control and achieves comparable performance to KPV and PMMR.

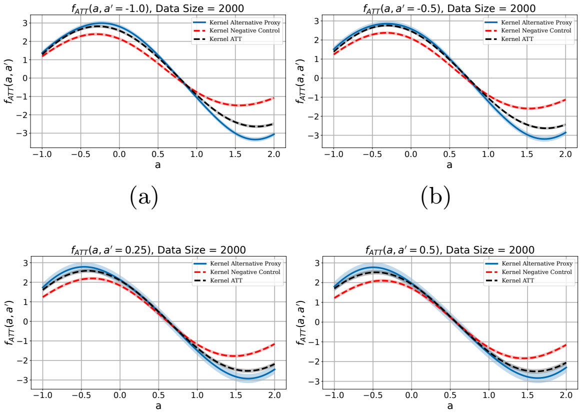
Legalized Abortion and Crime Dataset: We analyze the data from (Donohue and Levitt, 2001), as preprocessed by (Woody et al., 2020), following a similar approach to (Mastouri et al., 2021; Wu et al., 2024). The data is sourced from the GitHub repository of the code published by (Mastouri et al., 2021)[1]. The key variables in the causal graph are summarized as follows: (i) treatment variable (A): effective abortion rate, (ii) outcome varible (Y): murder rate, (iii) treatment proxy variable (Z): generosity to aid families with dependent children, and (iv) outcome proxy variable (W): beer consumption per capita, log-prisoner population per capita, and concealed weapons law. The remaining variables are captured by the unobserved confounding variable set U. Figure (2c) demonstrates the mean squared error results averaged over 30 realizations and compares our method to other approaches (we tested each of the 10 data files[1] with three different data splits). Our method outperforms PKIPW and delivers comparable results to other kernel-based methods. Compared to Kernel Negative Control, our method achieves a lower mean squared error but exhibits a higher variance.

Grade Retention and Cognitive Outcome: We apply our proposed method to study the effect of grade retention on the long-term cognitive outcome using data from the ECLS-K panel study (Fruehwirth et al., 2016; Deaner, 2023). We obtain the data from (Mastouri et al., 2021),[1] where the key variables are as follows: (i) treatment variable (A): grade retention, (ii) outcome variable (Y): cognitive test scores in Maths and Readings at age 11, (iii) treatment proxy variable (Z): the average of 1st/2nd and 3rd/4th year elementary scores, and (iv) outcome proxy variable (W): the cognitive and behavioral test scores from kindergarten. Figure (2d) presents the mean squared error results averaged over 30 realizations (3 realizations for each of the 10 data files), along with a comparison to other methods. In both of the datasets (IR: Reading grade retention; IM: Math grade retention), our proposed method performs better than PKIPW.

Further Comparison of Our Approach With Outcome Bridge-Based Methods: A key question as a result of our experiments is whether our method outperforms outcome bridge-based methods under certain conditions. To address this, we conduct an ablation study, detailed in S.M. (Sec. 13.5.1), comparing both approaches across six synthetic settings that vary the informativeness of two proxy variables, Z and W, relative to the confounders. The results, summarized in Table (1), indicate that our method performs better in settings where W is more informative, while outcome bridge-based methods excel when Z is more informative. Our analysis suggests that our method is more robust to violations of the completeness Assumption (3.7)—which ensures the existence of treatment bridge functions—while outcome bridge-based methods depend on this assumption for identifiability. We hypothesize that our method is more sensitive to violations of the identifiability condition (as defined by Assumption 3.3) than to violations of the bridge function existence condition. Further experiments with the Job Corps dataset in S.M. (Sec. 13.5.2) support these findings and highlight the complementary strengths of treatment and outcome bridge-based methods. Future work will explore these trade-offs in greater depth.

6.2 Synthetic Data Experiment for ATT

To demonstrate the effectiveness of our proposed method in conditional dose-response curve estimation, we use the low-dimensional synthetic data setting from the previous section. In particular, we train our method to estimate $f_{\mathrm{ATT}}(a, a')$ for different values of a'. Figure (3) shows dose-response estimates for $a' \in \{-1, -0.5, 0.25, 0.5\}$ using data size of 2000. We compare with Kernel-ATT (Singh et al., 2023) and Kernel Negative Control (Singh, 2023). The Kernel-ATT algorithm assumes access to the confounding variables U so we used the variables (A, Y, U) for this algorithm, making it an oracle method. Notably, our method produces results closer to Kernel-ATT than achieved by Kernel Negative Control. S.M. (Sec. 13.5.2) provides additional experimental results on the Job Corps dataset for conditional dose-response.



(a) (b)

(c) (d)

Figure 3: Conditional dose-response curve estimation for synthetic low-dimensional data across a' values and algorithms (averaged over 30 different runs) - mean solid line and standard deviation envelopes.

7 CONCLUSION

We propose a methodology for proxy causal learning that leverages a treatment bridge function. Our method enables the recovery of causal effects in the graphical model illustrated in Figure (1). It requires access to two proxy variables to the latent confounding variable, along with widely used completeness assumptions in the PCL setting. Our approach is practical for two key reasons. First, it avoids density ratio estimation - a challenging task in high dimensions. This enables strong performance even for high dimensional treatments, as demonstrated in the dSprite experiment. Second, the RKHS formulation of the problem allows for strong consistency guarantees, and closed form solutions that are easily implemented via matrix operations.

[1] `https://github.com/yuchen-zhu/kernel_proxies`

Acknowledgements

Bariscan Bozkurt, Dimitri Meunier, Liyuan Xu, and Arthur Gretton are supported by the Gatsby Charitable Foundation. We also thank the AISTATS 2025 reviewers for their valuable feedback and discussions.

References

Alabdulmohsin, I., Chiou, N., D'Amour, A., Gretton, A., Koyejo, S., Kusner, M. J., Pfohl, S. R., Salaudeen, O., Schrouff, J., and Tsai, K. (2023). Adapting to latent subgroup shifts via concepts and proxies. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9637–9661. PMLR.

Aubin, J.-P. (2011). *Applied functional analysis*. John Wiley & Sons.

Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962—-972.

Ben-Israel, A. and Greville, T. N. (2006). *Generalized inverses: theory and applications*. Springer Science & Business Media.

Blanchard, G. and Mücke, N. (2018). Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013.

Blundell, R., Horowitz, J., and Parey, M. (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics*, 3:29–51.

Buldygin, V. V. and Kozachenko, I. V. (2000). *Metric characterization of random variables and random processes*, volume 188. American Mathematical Soc.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368.

Choi, H. K., Hernán, M. A., Seeger, J. D., Robins, J. M., and Wolfe, F. (2002). Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *The Lancet*, 359(9313):1173–1177.

Connors, A., Speroff, T., Dawson, N., Thomas, C., Harrell, F., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, 276(11):889–897.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Deaner, B. (2023). Proxy controls and panel data.

Donohue, John J., I. and Levitt, S. D. (2001). The Impact of Legalized Abortion on Crime*. *The Quarterly Journal of Economics*, 116(2):379–420.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer Science & Business Media.

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

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *The Review of Economics and Statistics*, 94(1):153–171.

Fruehwirth, J. C., Navarro, S., and Takahashi, Y. (2016). How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects. *Journal of Labor Economics*, 34(4):979–1021.

Gretton, A. (2013). Introduction to RKHS, and some simple kernel algorithms. Advanced Topics in Machine Learning lecture, University College London.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012a). Conditional mean embeddings as regressors. In *International Conference on Machine Learningg.*

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. (2012b). Modelling transition dynamics in mdps with rkhs embeddings. In *International Conference on Machine Learning.*

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations.*

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning.*

Kallus, N., Mao, X., and Uehara, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.*, 10:1391–1445.

Klebanov, I., Schuster, I., and Sullivan, T. J. (2020). A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606.

Kompa, B., Bellamy, D., Kolokotrones, T., Beam, A., et al. (2022). Deep learning methods for proximal inference via maximum moment restriction. *Advances in Neural Information Processing Systems*.

Kress, R. (2013). *Linear Integral Equations*. Applied Mathematical Sciences. Springer New York.

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

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2022). Optimal rates for regularized conditional mean embedding learning. *Advances in Neural Information Processing Systems*.

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2024). Towards optimal sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25(181):1–51.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M. J., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Meanti, G., Carratino, L., De Vito, E., and Rosasco, L. (2022). Efficient hyperparameter tuning for large scale kernel ridge regression. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6554–6572. PMLR.

Meunier, D., Li, Z., Christensen, T., and Gretton, A. (2024). Nonparametric instrumental regression via kernel methods is minimax optimal. *arXiv preprint arXiv:2411.19653*.

Meunier, D., Li, Z., Gretton, A., and Kpotufe, S. (2023). Nonlinear meta-learning can guarantee faster rates. *arXiv preprint arXiv:2307.10870*.

Meunier, D., Shen, Z., Mollenhauer, M., Gretton, A., and Li, Z. (2025). Optimal rates for vector-valued spectral regularization learning algorithms. *Advances in Neural Information Processing Systems*, 37:82514–82559.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987—993.

Mollenhauer, M. and Koltai, P. (2020). Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*.

Mollenhauer, M., Mücke, N., and Sullivan, T. (2022). Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem. *arXiv preprint arXiv:2211.08875*.

Muandet, K., Jitkrittum, W., and Kübler, J. M. (2020). Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*.

Park, J. and Muandet, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems*.

Pearl, J. (2009). *Causality*. Cambridge University Press, 2 edition.

Pearl, J. and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence: Proceedings of the Eleventh Conference on Artificial Intelligence*, pages 444–453.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does job corps work? impact findings from the national job corps study. *American Economic Review*, 98(5):1864–86.

Singh, R. (2023). Kernel methods for unobserved confounding: Negative controls, proxies, and instruments.

Singh, R., Xu, L., and Gretton, A. (2023). Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In Hutter, M., Servedio, R. A., and Takimoto, E., editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg. Springer Berlin Heidelberg.

Song, L., Huang, J., Smola, A., and Fukumizu, K. (2009). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition.

Tsai, K., R Pfohl, S., Salaudeen, O., Chiou, N., Kusner, M., D'Amour, A., Koyejo, S., and Gretton, A. (2024). Proxy methods for domain adaptation. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3961–3969. PMLR.

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

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

Woody, S., Carvalho, C. M., Hahn, P. R., and Murray, J. S. (2020). Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime.

Wu, Y., Fu, Y., Wang, S., and Sun, X. (2024). Doubly robust proximal causal learning for continuous treatments. In *International Conference on Learning Representations*.

Xu, L. and Gretton, A. (2023). Causal benchmark based on disentangled image dataset.

Xu, L. and Gretton, A. (2024). Kernel single proxy control for deterministic confounding.

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

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*.

Checklist

1. For all models and algorithms presented, check if you include:
   - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]. We provide a detailed account of the necessary assumptions and mathematical settings for the algorithms presented in Sections (3) and (4). The derivations of the algorithms are elaborated in S.M. (Sec. 11).
   - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]. We include ablation studies, hyperparameter tuning procedures, and discussions on complexity in the S.M. (Sec. 13).
   - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]. We provide an anonymous GitHub URL for our implementation in the S.M. (Sec. 6). (For the camera-ready version of the paper, we include a non-anonymized GitHub URL in S.M.(Sec. 13.4).)

2. For any theoretical claim, check if you include:
   - (a) Statements of the full set of assumptions of all theoretical results. [Yes].
   - (b) Complete proofs of all theoretical results. [Yes]. For identifiability and existence proofs, see the S.M. (Sec. 9) and (Sec. 10). Consistency proofs are provided S.M. (12). For algorithm derivations, refer to the S.M. (13)
   - (c) Clear explanations of any assumptions. [Yes].

3. For all figures and tables that present empirical results, check if you include:
   - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]. See the anonymous GitHub URL for the implementation code, which includes a README.md file with instructions for reproducing the results.
   - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]. Refer to the S.M. (13) for hyperparameter tuning procedures, ablation studies, and further details on numerical experiments.
   - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]. Our experimental results are based on multiple realizations, with box plots and/or standard deviation envelopes illustrating the variability.
   - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]. We did not include details on computational resources since our method does not require advanced computing infrastructure. Each causal learning experiment can be run on a standard computer.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - (a) Citations of the creator If your work uses existing assets. [Yes]. For 'Legalized Abortion and Crime Dataset', we cite (Donohue and Levitt, 2001; Woody et al., 2020; Mastouri et al., 2021) in Section (6). For 'Grade Retention Experiment', we cite (Fruehwirth et al., 2016; Deaner, 2023) in Section (6). The other experiments are based on synthetic data generation processes.
   - (b) The license information of the assets, if applicable. [Not Applicable].
   - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   - (d) Information about consent from data providers/curators. [Not Applicable]
   - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for Density Ratio-based Proxy Causal Learning Without Density Ratios

Section (8) reviews the previous work on treatment bridge function-based and outcome bridge function-based proxy causal learning frameworks. In Section (9), we provide the proofs for identifiability of the dose-response and the conditional dose-response curves. Section (10) contains the proofs for existence of the bridge function. In Section (11), we derive the algorithms for estimating the dose-response and the conditional dose-response curves. The proofs for the consistency results of our approach are provided in Section (12). Section (13) offers the additional details on the numerical experiments and ablation studies while Section (14) discusses the identifiability of the dose-response in the discrete case as a specific example. Finally, Section (15) discusses the extension of our method to settings with additional observed confounding variables.

8 COMPARISON WITH OTHER METHODS

In Section (6) and S.M. (Sec. 13.5), we compare our method with other proxy causal learning algorithms: Proximal Kernel Inverse Probability Weighted (PKIPW) (Wu et al., 2024), Kernel Proxy Variable (KPV) (Mastouri et al., 2021), Proximal Maximum Moment Restriction (PMMR) (Mastouri et al., 2021), and Kernel Negative Control (KNC) (Singh, 2023). Here, we further examine these baselines, focusing on their assumptions, bridge functions, and suitability for high-dimensional settings.

Proximal Kernel Inverse Probability Weighted (PKIPW): PKIPW extends the binary treatment framework of (Cui et al., 2024) to continuous treatments (Wu et al., 2024). Cui et al. (2024) introduced a treatment bridge function $\varphi_0(z, a)$ satisfying

$$\mathbb{E}[\varphi_0(Z, a)|W, A = a] = \frac{1}{p(A = a|W)}$$

to identify the ATE in the binary case via

$$f_{ATE}(a) = \mathbb{E}[\mathbf{1}[A = a]Y\varphi_0(Z, A)].$$

This result relies on the following completeness assumptions (Assumptions (10) and (11) in (Cui et al., 2024)):

Assumption 8.1. *Let $\ell_1 : \mathcal{U} \to \mathbb{R}$ and $\ell_2 : \mathcal{W} \to \mathbb{R}$ be square integrable functions. Assume that the following conditions hold for all $a \in \mathcal{A}$,:*

i. $\mathbb{E}[\ell_1(U)|W = w, A = a] = 0 \quad \forall w \in \mathcal{W}$ *if and only if* $\ell_1(U) = 0 \quad p(U)-$*almost everywhere*

ii. $\mathbb{E}[\ell_2(W)|Z = z, A = a] = 0 \quad \forall z \in \mathcal{Z}$ *if and only if* $\ell_2(W) = 0 \quad p(W)-$*almost everywhere*

Furthermore, they leverage the following completeness assumption in order to obtain the uniqueness of the treatment bridge function:

Assumption 8.2. *Let $\ell : \mathcal{Z} \to \mathbb{R}$ be any square integrable function. Assume that the following conditions hold for all $a \in \mathcal{A}$:* $\mathbb{E}[\ell(Z)|W = w, A = a] = 0 \quad \forall w \in \mathcal{W}$ *if and only if* $\ell(Z) = 0 \quad p(Z)-$*almost everywhere.*

PKIPW adapts this to continuous treatments by replacing the indicator function with a kernel function:

$$f_{ATE}(a) = \mathbb{E}[\mathbf{1}[A = a]Y\varphi_0(Z, A)] \approx \frac{1}{n}\sum_{i=1}^{n} K_l(a_i - a)\varphi_0(z_i, a)y_i, \tag{4}$$

where K_l is a kernel function with the bandwidth variable l. In particular, their approximation relies on their result

$$\mathbb{E}[\mathbf{1}[A = a]Y\varphi_0(Z, A)] = \lim_{l \to 0}[K_l(A - a)Y\varphi_0(Z, A)]$$

under the condition that $\mathbb{E}[\mathbf{1}[A=a]Y\varphi_0(Z,A)]$ is continuous and bounded uniformly with respect to a (see Theorem 4.2 in (Wu et al., 2024)).

PKIPW requires estimating the policy function $\frac{1}{p(A=a|W)}$ via kernel density estimation or conditional normalizing flows (CNFs) which adds computational complexity. In contrast, our approach bypasses density ratio estimation by modifying the bridge function (see Section (4)). Moreover, while PKIPW assumes uniqueness of the treatment bridge function with Assumption (8.2), we show that convergence to the minimum RKHS norm solution suffices for consistency. Additionally, PKIPW's completeness condition on Z (Assumption (8.1-ii)) is stronger than our Assumption (3.7). Thus, our approach is more feasible and computationally desirable as reinforced by the experimental results shown in Figures (2a), (2c), and (2d). In Figure (2b) for dSprite experiment with high dimensional treatment, we could not produce result for PKIPW due to its published code being limited to univariate treatments. However, our method successfully handles high-dimensional treatments, as evidenced by the results in the dSprite experiment, showcasing another advantage.

Outcome Bridge Function-Based Approaches with Kernel Methods: (Mastouri et al., 2021) and (Singh, 2023) introduced kernel-based PCL approaches leveraging an outcome bridge function $h(w,a)$ that satisfies

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

Given this outcome bridge function, Wang Miao and Tchetgen (2024) showed that the dose-response can be identified as

$$f_{ATE}(a)=\int h(a,w)p(w)dw.$$

Both (Mastouri et al., 2021) and (Singh, 2023) rely on the following completeness conditions:

Assumption 8.3. *Let $\ell_1:\mathcal{U}\to\mathbb{R}$ and $\ell_2:\mathcal{W}\to\mathbb{R}$ be square integrable functions. Assume that the following conditions hold for all $a\in\mathcal{A}$, $x\in\mathcal{X}$:*

i. $\mathbb{E}[\ell_1(U)|Z=z,A=a]=0\quad\forall z\in\mathcal{Z}$ *if and only if* $\ell_1(U)=0\quad p(U)-$*almost everywhere*

ii. $\mathbb{E}[\ell_2(Z)|W=w,A=a]=0\quad\forall w\in\mathcal{W}$ *if and only if* $\ell_2(Z)=0\quad p(Z)-$*almost everywhere.*

(Xu et al., 2021; Deaner, 2023) later showed that Assumption (8.3-ii) can be replaced by the weaker condition:

Assumption 8.4. *Let $\ell:\mathcal{U}\to\mathbb{R}$ be any square integrable function. Assume that the following conditions hold for all $a\in\mathcal{A}$:* $\mathbb{E}[\ell(U)|W=w,A=a]=0\quad\forall w\in\mathcal{W}$ *if and only if* $\ell(U)=0\quad p(U)-$*almost everywhere.*

Notably, these completeness conditions align with those in our approach: Assumption (3.3) corresponds to Assumption (8.4), while Assumption (3.7) matches Assumption (8.3-i). However, they serve different purposes. We use Assumption (3.3) to ensure identifiability of the causal structural function with treatment bridge function, while outcome bridge-based methods use it to guarantee the existence of the outcome bridge function. Similarly, we employ Assumption (3.7) to establish the existence of the treatment bridge function, whereas outcome bridge-based methods use it for causal function identifiability.

Mastouri et al. (2021) introduced two estimation algorithms: KPV, which learns the bridge function in two stages, and PMMR, which leverages a closed-form solution under the Maximum Moment Restriction framework (Muandet et al., 2020). Singh (2023) proposed KNC, an alternative kernel-based approach with a different outcome bridge function representation. Since both methods employ kernel techniques, they are well-suited for high-dimensional settings. For algorithmic details, we refer readers to the respective papers.

9 IDENTIFICATION PROOFS

9.1 Identifiability of Dose-Response Curve

In this section, we prove Theorem (3.4-1.) for f_{ATE}. For completeness, we state the theorem here again.

Theorem 9.1. *Suppose that there exists a square integrable bridge function $\varphi_0 : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ that satisfies the following functional equation*

$$\mathbb{E}[\varphi_0(Z,a)|W, A=a] = \frac{p(W)p(A)}{p(W,A)}, \quad a \in \mathcal{A}.$$

Furthermore, suppose that Assumptions (3.2) and (3.3) hold. Then, the dose-response curve is given by

$$f_{ATE}(a) = \mathbb{E}[Y\varphi_0(Z,a)|A=a], \quad a \in \mathcal{A}.$$

Proof. Suppose that

$$\mathbb{E}[\varphi_0(Z,A)|W,A] = \frac{p(W)p(A)}{p(W,A)}.$$

Then, note the following, for $a \in \mathcal{A}$,

$$\begin{aligned}\mathbb{E}[\varphi_0(Z,a)|W,A=a] &= \mathbb{E}_{U|W,A=a}[\mathbb{E}[\varphi_0(Z,a)|U,W,A=a]] \quad \text{(by Law of Total Expectation)}\\ &= \mathbb{E}_{U|W,A=a}[\mathbb{E}[\varphi_0(Z,a)|U,A=a]] \quad \text{(since } Z \perp W|U\text{, Assumption (3.2))}\end{aligned}$$

On the other hand,

$$\begin{aligned}p(w) &= \int p(w|u)p(u)du = \int p(w|a,u)p(u)du \quad \text{(since } W \perp A|U\text{, Assumption (3.2))}\\ &= \int \frac{p(u|w,a)p(w|a)}{p(u|a)}p(u)du \quad \text{(Baye's Rule)}\\ &= \int \frac{p(u|w,a)p(w,a)}{p(u,a)}p(u)du = p(w,a)\int \frac{p(u)}{p(u,a)}p(u|w,a)du\\ &= p(w,a)\mathbb{E}\left[\frac{p(U)}{p(U,a)}\middle|W=w,A=a\right]\end{aligned}$$

As a result,

$$\frac{p(W)}{p(W,a)} = \mathbb{E}\left[\frac{p(U)}{p(U,a)}\middle|W,A=a\right] \Rightarrow \frac{p(W)p(a)}{p(W,a)} = \mathbb{E}\left[\frac{p(U)p(a)}{p(U,a)}\middle|W,A=a\right]. \tag{5}$$

For every value $a \in \mathcal{A}$, it therefore holds $p(W)$-a.e. that

$$\mathbb{E}_{U|W,A=a}[\mathbb{E}[\varphi_0(Z,a)|U,A=a]] = \mathbb{E}\left[\frac{p(U)p(a)}{p(U,a)}\middle|W,A=a\right].$$

Therefore, due to Assumption (3.3), we have for all $a \in \mathcal{A}$,

$$\mathbb{E}[\varphi_0(Z,a)|U,A=a] = \frac{p(U)p(a)}{p(U,a)} \quad p(U)-\text{a.e.} \tag{6}$$

Next, we observe that

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[Y|A=a,U]] = \int \mathbb{E}[Y|A=a,u]p(u)du = \int \mathbb{E}[Y|A=a,u]\frac{p(u)p(a)}{p(u,a)}\frac{p(u,a)}{p(a)}du\\ &= \int \mathbb{E}[Y|A=a,u]\frac{p(u)p(a)}{p(u,a)}p(u|a)du = \mathbb{E}_{U|A=a}\left[\mathbb{E}[Y|A=a,U]\frac{p(U)p(a)}{p(U,a)}\right]\\ &= \mathbb{E}_{U|A=a}\left[\mathbb{E}[Y|A=a,U]\mathbb{E}[\varphi_0(Z,a)|U,A=a]\right] \quad \text{(by Equation 6)}\\ &= \mathbb{E}_{U|A=a}\left[\int yp(y|a,U)dy\int \varphi_0(z,a)p(z|U,a)dz\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{U|A=a}\left[\int \varphi_0(z,a)\left(\int yp(y|a,U)dy\right)p(z|U,a)dz\right] \\
&= \mathbb{E}_{U|A=a}\left[\int \varphi_0(z,a)\left(\int yp(y|a,U,z)dy\right)p(z|U,a)dz\right] \quad (\text{since } Y \perp Z|A,U, \text{ Assump. (3.2)}) \\
&= \mathbb{E}_{U|A=a}\left[\int\int \varphi_0(z,a)y\underbrace{p(y|a,U,z)p(z|U,a)}_{p(y,z|a,U)}dydz\right] \\
&= \int\int\int \varphi_0(z,a)y\underbrace{p(y,z|a,u)p(u|a)}_{p(u,y,z|a)}dydzdu \\
&= \int\int \varphi_0(z,a)y\underbrace{\int p(u,y,z|a)du}_{p(y,z|a)}dydz \\
&= \int\int \varphi_0(z,a)yp(y,z|a)dydz = \mathbb{E}[Y\varphi_0(Z,a)|A=a].
\end{aligned}
$$

As a result, we obtain

$$\mathbb{E}[Y\varphi_0(Z,a)|A=a] = \mathbb{E}[\mathbb{E}[Y|A=a,U]],$$

which indicates that $f_{\mathrm{ATE}}(a) = \mathbb{E}[Y\varphi_0(Z,a)|A=a]$. □

9.2 Identifiability of Conditional Dose-Response Curve

In this section, we prove Theorem (3.4-2.) for f_{ATT}. For completeness, we state the theorem here again.

Theorem 9.2. *Suppose there exists a square integrable bridge function $\varphi_0 : \mathcal{Z}\times\mathcal{A}\times\mathcal{A}\to\mathbb{R}$ that satisfies the following functional equation*

$$\mathbb{E}[\varphi_0(Z,a,a')|W,A=a] = \frac{p(W,a')p(a)}{p(W,a)p(a')}, \qquad (a,a')\in\mathcal{A}^2.$$

Furthermore, suppose that Assumptions (3.2) and (3.3) hold. Then, the conditional dose-response is given by

$$f_{ATT}(a,a') = \mathbb{E}[Y\varphi_0(Z,a,a')|A=a], \qquad (a,a')\in\mathcal{A}^2.$$

Proof. First, observe the following, for $(a,a')\in\mathcal{A}^2$,

$$
\begin{aligned}
&\mathbb{E}[\varphi_0(Z,a,a')|W,A=a] = \mathbb{E}_{U|W,A=a}\big[\mathbb{E}[\varphi_0(Z,a,a')|U,W,A=a]\big] \quad (\text{by Law of Total Expectation}) \\
&= \mathbb{E}_{U|W,A=a}\big[\mathbb{E}[\varphi_0(Z,a,a')|U,A=a]\big] \quad (\text{since } Z\perp W|U,A=a, \text{ Assumption (3.2)}) \qquad (7)
\end{aligned}
$$

Furthermore, note that

$$
\begin{aligned}
p(w,a') &= \int p(w,a'|u)p(u)du = \int p(w|u)p(a'|u)p(u)du \quad (\text{since } W\perp A|U, \text{ Assumption (3.2)}) \\
&= \int p(w|a,u)p(a'|u)p(u)du \quad (\text{again due to } W\perp A|U, \text{ Assumption (3.2)}) \\
&= \int \frac{p(u|w,a)p(w|a)}{p(u|a)}p(a'|u)p(u)du \quad (\text{Baye's Rule}) \\
&= \int \frac{p(u|w,a)p(w,a)}{p(u,a)}p(u,a')du = p(w,a)\int\frac{p(u,a')}{p(u,a)}p(u|w,a)du.
\end{aligned}
$$

As a result,

$$\frac{p(W,a')}{p(W,a)} = \mathbb{E}\left[\frac{p(U,a')}{p(U,a)}\middle|W,A=a\right] \Rightarrow \frac{p(W,a')p(a)}{p(W,a)p(a')} = \mathbb{E}\left[\frac{p(U,a')p(a)}{p(U,a)p(a')}\middle|W,A=a\right]. \qquad (8)$$

Recall that our assumption was

$$\mathbb{E}[\varphi_0(Z,a,a')|W,A=a] = \frac{p(W,a')p(a)}{p(W,a)p(a)}.$$

Combining Equation (7) and (8), for every value $a, a' \in \mathcal{A}$, it therefore holds $p(W)$-a.e. that

$$\mathbb{E}_{U|W,A=a}\Big[\mathbb{E}[\varphi_0(Z,a,a')|U,A=a]\Big] = \mathbb{E}_{U|W,A=a}\left[\frac{p(U,a')p(a)}{p(U,a)p(a')}\right].$$

Using the completeness Assumption (3.3), we obtain, for all $a, a' \in \mathcal{A}$

$$\mathbb{E}[\varphi_0(Z,a,a')|U,A=a] = \frac{p(U,a')p(a)}{p(U,a)p(a')} \quad p(U)\text{-a.e.} \tag{9}$$

Next, to obtain the ATT function, consider

$$\begin{aligned}
&f_{\mathrm{ATT}}(a,a') = \mathbb{E}_{U|A=a'}[\mathbb{E}[Y|A=a,U]] = \int \mathbb{E}[Y|A=a,u]p(u|a')du \\
&= \int \mathbb{E}[Y|A=a,u]\frac{p(u,a')}{p(a')}\frac{p(a)}{p(u,a)}\frac{p(u,a)}{p(a)}du \\
&= \int \mathbb{E}[Y|A=a,u]\frac{p(u,a')p(a)}{p(u,a)p(a')}p(u|a)du \\
&= \mathbb{E}_{U|A=a}\Big[\mathbb{E}[Y|A=a,U]\mathbb{E}[\varphi_0(Z,a,a')|U,A=a]\Big] \quad \text{(by Equation (9))} \\
&= \mathbb{E}_{U|A=a}\left[\int yp(y|a,U)dy \int \varphi_0(z,a,a')p(z|a,U)dz\right] \\
&= \mathbb{E}_{U|A=a}\left[\int\int y\varphi_0(z,a,a')p(y|a,U)p(z|a,U)dydz\right] \\
&= \mathbb{E}_{U|A=a}\left[\int\int y\varphi_0(z,a,a')\underbrace{p(y|a,U,z)p(z|a,U)}_{p(y,z|a,U)}dydz\right] \quad (Y \perp Z|U,A,\ \text{Assumption (3.2) }) \\
&= \mathbb{E}_{U|A=a}\left[\int\int y\varphi_0(z,a,a')p(y,z|a,U)dydz\right] \\
&= \int\int\int y\varphi_0(z,a,a')\underbrace{p(y,z|a,u)p(u|a)}_{p(u,y,z|a)}dydzdu \\
&= \int\int y\varphi_0(z,a,a')\underbrace{\int p(u,y,z|a)du}_{p(y,z|a)}dydz \\
&= \int\int y\varphi_0(z,a,a')p(y,z|a)dydz = \mathbb{E}[Y\varphi_0(Z,a,a')|A=a].
\end{aligned}$$

Hence, we have shown that $f_{\mathrm{ATT}}(a,a') = \mathbb{E}_{U|A=a'}[\mathbb{E}[Y|A=a,U]] = \mathbb{E}[Y\varphi_0(Z,a,a')|A=a]$. □

10 EXISTENCE OF BRIDGE FUNCTIONS

Our proofs follow the strategy of (Deaner, 2023). We note that Assumption (3.7) is weaker that the assumption used in (Cui et al., 2024; Wu et al., 2024). Namely they use the stronger assumption that for any square integrable function $\ell : \mathcal{W} \to \mathbb{R}$, for all $a \in \mathcal{A}$: $\mathbb{E}[\ell(W)|Z,A=a] = 0$ $p(Z)-$a.e. if and only if $\ell(W) = 0$ $p(W)$-a.e.

10.1 Existence of Bridge Function for Dose-Response Curve

We will discuss conditions that will guarantee the existence of the bridge function φ_0 for the functional Equation (1). To this end, we consider the following conditional mean operator

$$E_a : \mathcal{L}_2(\mathcal{W}, p_{W|A=a}) \to \mathcal{L}_2(\mathcal{Z}, p_{Z|A=a}),$$

such that

$$E_a \ell(z) = \mathbb{E}[\ell(W)|Z = z, A = a], \qquad z \in \mathcal{Z}$$

whose adjoint is given by

$$E_a^* : \mathcal{L}_2(\mathcal{Z}, p_{Z|A=a}) \to \mathcal{L}_2(\mathcal{W}, p_{W|A=a})$$

such that

$$E_a^* \ell(w) = \mathbb{E}[\ell(Z)|W = w, A = a], \qquad w \in \mathcal{W},$$

where $\mathcal{L}_2(\mathcal{W}, p_{W|A=a})$ denotes the square integrable functions of $w \in \mathcal{W}$ with respect to the distribution $p(W|A = a)$ and $\mathcal{L}_2(\mathcal{Z}, p_{Z|A=a})$ denotes the square integrable functions of $z \in \mathcal{Z}$ with respect to the distribution $p(Z|A = a)$. Our result will rely on Picard's Theorem as stated below.

Theorem 10.1 (Picard's Theorem; Theorem 15.8 in (Kress, 2013))**.** *Let $\mathcal{X}, \mathcal{Y}$ be Hilbert spaces and let $A : \mathcal{X} \to \mathcal{Y}$ be a compact linear operator with singular system $(\mu_n, \varphi_n, g_n)_{n=1}^{\infty}$. The equation of the first kind*

$$A\varphi = f$$

admits a solution if and only if $f \in \text{null}(A^*)^{\perp}$ and

$$\sum_{n=1}^{\infty} \frac{1}{\mu_n^2} |\langle f, g_n \rangle_{\mathcal{Y}}|^2 < \infty.$$

Here, A^* is the adjoint of the operator A and $\text{null}(A^*)$ represents the null-space of the operator A^*. Then a solution is given by

$$\varphi = \sum_{n=1}^{\infty} \frac{1}{\mu_n} \langle f, g_n \rangle_{\mathcal{Y}} \varphi_n.$$

Lemma 10.2. *Suppose Assumptions (3.2) and (3.7) hold. Then,*

$$\text{null}(E_a) \subseteq \big\{\ell \in \mathcal{L}_2(\mathcal{W}, p_{W|A=a}) \;\big|\; \mathbb{E}[\ell(W)|U, A = a] = 0\big\}.$$

Proof. Observe that $\text{null}(E_a) = \big\{\ell \in \mathcal{L}_2(\mathcal{W}, p_{W|A=a}) \;\big|\; \mathbb{E}[\ell(W)|Z, A = a] = 0\big\}$. Let $\ell \in \text{null}(E_a)$, then,

$$\begin{aligned} 0 &= E_a \ell = \mathbb{E}[\ell(W)|Z = \cdot, A = a] \\ &= \mathbb{E}_{U|Z=\cdot,A=a}\big[\mathbb{E}[\ell(W)|U, Z = \cdot, A = a]\big] \\ &= \mathbb{E}_{U|Z=\cdot,A=a}\big[\mathbb{E}[\ell(W)|U, A = a]\big] \quad \text{(since } W \perp Z|U, A\text{, Assumption (3.2))} \end{aligned}$$

Assumption (3.7) implies that $\mathbb{E}[\ell(W)|U, A = a] = 0$ almost surely. Hence,

$$\ell \in \big\{\ell \in \mathcal{L}_2(\mathcal{W}, p_{W|A=a}) \;\big|\; \mathbb{E}[\ell(W)|U, A = a] = 0\big\}.$$

□

Let us introduce the notation $\varphi_0^a := \varphi_0(\cdot, a)$ and $r_a(W) := \frac{p(W)p(a)}{p(W,a)}$. The bridge function is then solution of

$$E_a^* \varphi_0^a = r_a.$$

E_a is well-defined, linear and bounded with $\|E_a\| \leq 1$. In order to apply Theorem (10.1), we make the following additional assumptions.

Assumption 10.3. *For each $a \in \mathcal{A}$, the operator E_a^* is compact with singular system $\{\mu_{a,n}, \varphi_{a,n}, g_{a,n}\}_{n=1}^{\infty}$.*

Assumption 10.4. *The density ratio function r_a satisfies*

$$\sum_{n=1}^{\infty} \frac{1}{\mu_{a,n}^2} \left| \langle r_a, g_{a,n} \rangle_{\mathcal{L}_2(\mathcal{W}, p_{W|A=a})} \right|^2 < \infty.$$

Lemma 10.5. *Suppose that Assumptions (3.2) and (3.7) hold. Then, $r_a \in \operatorname{null}(E_a)^{\perp}$.*

Proof. Recall that we previously proved that (Eq. 5),

$$\frac{p(W)p(a)}{p(W,a)} = \mathbb{E}\Big[\frac{p(U)p(a)}{p(U,a)}\Big| W, A = a\Big].$$

Let $\ell \in \operatorname{null}(E_a)$. Then,

$$\begin{aligned}
\big\langle \ell, r_a \big\rangle_{\mathcal{L}_2(\mathcal{W}, p_{W|A=a})} &= \mathbb{E}[\ell(W) r_a(W) | A = a] \\
&= \mathbb{E}\Big[\ell(W)\mathbb{E}\Big[\frac{p(U)p(a)}{p(U,a)}\Big| W, A = a\Big]\Big| A = a\Big] \\
&= \iint \ell(w) \frac{p(u)p(a)}{p(u,a)} p(u|w,a) p(w|a) du dw \\
&= \iint \ell(w) \frac{p(u)p(a)}{p(u,a)} p(u,w|a) du dw \\
&= \iint \ell(w) \frac{p(u)p(a)}{p(u,a)} p(w|u,a) p(u|a) du dw \\
&= \mathbb{E}\Big[\mathbb{E}[\ell(W)|U, A = a] \frac{p(U)p(a)}{p(U,a)}\Big| A = a\Big] = 0,
\end{aligned}$$

where the last equality is due to Lemma (10.2). Therefore, $r_a \in \operatorname{null}(E_a)^{\perp}$. □

Theorem 10.6. *Suppose that Assumptions (3.2), (3.7), (10.3), and (10.4) hold. Then, there exists a solution to the functional equation*

$$\mathbb{E}[\varphi_0(Z, a) | W, A = a] = \frac{p(W)p(a)}{p(W,a)}, \qquad a \in \mathcal{A}.$$

Proof. Fix $a \in \mathcal{A}$. The functional equation can be written as $E_a^* \varphi_0^a = r_a$. In Lemma (10.5), we proved that $r_a \in \operatorname{null}(E_a)^{\perp}$. Furthermore, combined with Assumptions (10.3) and (10.4), Theorem (10.1) implies the existence. □

10.2 Existence of the Bridge Function for Conditional Dose-Response

Let us fix $a, a' \in \mathcal{A}$. In this section, we introduce the notations:

$$\begin{aligned}
\varphi_0^{a,a'} &= \varphi_0(\cdot, a, a'), \\
r_{a,a'}(W) &= \frac{p(W, a')p(a)}{p(W,a)p(a')}.
\end{aligned}$$

Then, the functional equation of the ATT bridge function

$$\mathbb{E}[\varphi_0(Z,a,a')|W,A=a] = \frac{p(W,a')p(a)}{p(W,a)p(a')},$$

can be written as $E_a^* \varphi_0^{a,a'} = r_{a,a'}$, where E_a^* is the conditional expectation operator that we defined in S.M. (10.1). Our result will again rely on Theorem (10.1), so we make the following assumption.

Assumption 10.7. *The density ratio function $r_{a,a'}$ satisfies*

$$\sum_{n=1}^{\infty} \frac{1}{\mu_{a,n}^2} \left| \langle r_{a,a'}, g_{a,n} \rangle_{\mathcal{L}_2(\mathcal{W}, p_{W|A=a})} \right|^2 < \infty.$$

Lemma 10.8. *Suppose that Assumptions (3.2) and (3.7) hold. Then, $r_{a,a'} \in \mathrm{null}(E_a)^{\perp}$.*

Proof. Recall that we previously proved that (Eq. (8),

$$\frac{p(W,a')p(a)}{p(W,a)p(a')} = \mathbb{E}\Big[\frac{p(U,a')p(a)}{p(U,a)p(a')}\Big| W, A=a\Big].$$

Let $\ell \in \mathrm{null}(E_a)$. Then,

$$\begin{aligned}
\big\langle \ell, r_{a,a'} \big\rangle_{\mathcal{L}_2(\mathcal{W}, p_{W|A=a})} &= \mathbb{E}[\ell(W) r_{a,a'}(W)|A=a] \\
&= \mathbb{E}\Big[\ell(W)\mathbb{E}\Big[\frac{p(U,a')p(a)}{p(U,a)p(a')}\Big| W, A=a\Big]\Big| A=a\Big] \\
&= \iint \ell(w) \frac{p(u,a')p(a)}{p(u,a)p(a')} p(u|w,a)p(w|a)dudw \\
&= \iint \ell(w) \frac{p(u,a')p(a)}{p(u,a)p(a')} p(u,w|a)dudw \\
&= \iint \ell(w) \frac{p(u,a')p(a)}{p(u,a)p(a')} p(w|u,a)p(u|a)dudw \\
&= \mathbb{E}\Big[\mathbb{E}[\ell(W)|U,A=a]\frac{p(U,a')p(a)}{p(U,a)p(a')}\Big| A=a\Big] = 0,
\end{aligned}$$

where the last equality is due to Lemma (10.2). Therefore, $r_{a,a'} \in \mathrm{null}(E_a)^{\perp}$. □

Theorem 10.9. *Suppose that Assumptions (3.2), (3.7), (10.3), and (10.7) hold. Then, there exists a solution to the functional equation*

$$\mathbb{E}[\varphi_0(Z,a,a')|W,A=a] = \frac{p(W,a')p(a)}{p(W,a)p(a')}.$$

Proof. Fix $a, a' \in \mathcal{A}$. The functional equation can be written as $E_a^* \varphi_0^{a,a'} = r_{a,a'}$. In Lemma (10.8), we proved that $r_{a,a'} \in \mathrm{null}(E_a)^{\perp}$. Furthermore, combined with Assumptions (10.3) and (10.7), Theorem (10.1) implies the existence. □

11 ALGORITHM DERIVATIONS

11.1 Dose-Response Curve Algorithm

Here, we will derive the Algorithm (4.1). For completeness, we restate the algorithm first.

Algorithm (Algorithm (4.1)). *Let $\{z_i, w_i, a_i\}_{i=1}^n$ and $\{\tilde{w}_i, \tilde{a}_i\}_{i=1}^m$ be the first-stage and second-stage data, respectively, and $(\lambda_1, \lambda_2, \lambda_3)$ be the regularization parameters of first, second, and third stage regressions. Furthermore, let $\{\alpha_i\}_{i=1}^{m+1}$ be the minimizer of the following cost function:*

$$\hat{\mathcal{L}}^{2SR}(\boldsymbol{\alpha}) = \frac{1}{m} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\mathbf{1}}{m})^T \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big] \frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}$$

$$
\begin{aligned}
&- 2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T \Big[\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] \frac{\mathbf{1}}{m} \end{bmatrix} \\
&+ \lambda_2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]^T & (\frac{\mathbf{1}}{m})^T \Big[\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] \frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}
\end{aligned}
$$

where $\boldsymbol{I} \in \mathbb{R}^{n \times n}$ is the identity matrix, $\mathbf{1}$ is vector of ones, $\alpha_{1:m} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_m \end{bmatrix}^T \in \mathbb{R}^m$, $\boldsymbol{B} = (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{W}} \odot \boldsymbol{K}_{A\tilde{A}}) \in \mathbb{R}^{n \times m}$, and $\bar{\boldsymbol{B}}$ is the matrix whose j-th column is given by

$$
\bar{\boldsymbol{B}}_{:,j} = \frac{1}{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j}).
$$

Then, the dose-response curve estimation $f_{ATE}(a)$ is given by

$$
\begin{aligned}
f_{ATE}(a) = \alpha_{1:m}^T &\Big(\boldsymbol{B}^T \big(\boldsymbol{K}_{ZZ} \mathit{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} \big) \odot \boldsymbol{K}_{\tilde{A}a} \Big) \\
&+ \alpha_{m+1} \Big(\bar{\boldsymbol{B}}^T \big(\boldsymbol{K}_{ZZ} \mathit{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} \big) \odot \boldsymbol{K}_{\tilde{A}a} \Big) \frac{\mathbf{1}}{m}
\end{aligned}
$$

Here, for $F \in \{A, W, Z\}$ with domain $\mathcal{F}$, the first-stage kernel matrices are denoted as $\boldsymbol{K}_{FF} = [k_{\mathcal{F}}(f_i, f_j)]_{ij} \in \mathbb{R}^{n \times n}$, $\boldsymbol{K}_{Ff} = [k_{\mathcal{F}}(f_i, f)]_i \in \mathbb{R}^n$, where $\{f_i\}_{i=1}^n$ denotes the first-stage data samples. Similarly, with second-stage variables $\tilde{F} \in \{\tilde{A}, \tilde{W}\}$, the kernel matrices are denoted as follows: $\boldsymbol{K}_{\tilde{F}\tilde{F}} = [k_{\mathcal{F}}(\tilde{f}_i, \tilde{f}_j)]_{ij} \in \mathbb{R}^{m \times m}$, $\boldsymbol{K}_{F\tilde{F}} = [k_{\mathcal{F}}(f_i, \tilde{f}_j)]_{ij} \in \mathbb{R}^{n \times m}$, $\boldsymbol{K}_{F\tilde{f}} = [k_{\mathcal{F}}(f_i, \tilde{f})]_i \in \mathbb{R}^n$, and $\boldsymbol{K}_{\tilde{F}f} = [k_{\mathcal{F}}(\tilde{f}_j, f)]_j \in \mathbb{R}^m$.

Derivation of Algorithm (4.1). Let $r(W, a)$ denote $p(W)p(a)/p(W, a)$. We would like to find the optimum of the following loss function

$$
\mathcal{L}^{2SR}(\varphi) = \mathbb{E}_{W,A}[(r(W, A) - \mathbb{E}[\varphi(Z, A)|W, A])^2] + \lambda_2 \|\varphi_0\|^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}}.
$$

This loss cannot be directly optimized as it involves the conditional mean $\mathbb{E}[\varphi(Z, A)|W, A]$. A similar problem in (Mastouri et al., 2021; Singh, 2023; Xu and Gretton, 2024) is addressed using a two-stage regression approach: the first stage approximates the conditional expectation, and the second stage minimizes the loss. We build on the approach of Xu and Gretton (2024) in our algorithm derivation, which is shown to be more numerically stable, and to require optimizing fewer parameters, than the earlier approaches: see Xu and Gretton (2024, Appendix F)

First Stage: Assume that the bridge function φ is the RKHS $\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}$. Then,

$$
\begin{aligned}
\mathbb{E}[\varphi(Z, a)|W = w, A = a] &= \mathbb{E}[\langle \varphi, \phi_{\mathcal{Z}}(Z) \otimes \phi_{\mathcal{A}}(a) \rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} | W = w, A = a] \\
&= \langle \varphi, \mathbb{E}[\phi_{\mathcal{Z}}(Z)|W = w, A = a] \otimes \phi_{\mathcal{A}}(a) \rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \langle \varphi, \mu_{Z|W,A}(w, a) \otimes \phi_{\mathcal{A}}(a) \rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}}
\end{aligned}
$$

where $\mu_{Z|W,A}(w, a) = \mathbb{E}[\phi_{\mathcal{Z}}(Z)|W = w, A = a]$ is the CME. Considering the sample-based first-stage regression with given data $\{z_i, w_i, a_i\}_{i=1}^n$

$$
\hat{\mathcal{L}}^c(C) = \frac{1}{n} \sum_{i=1}^{n} \|\phi_{\mathcal{Z}}(z_i) - C(\phi_{\mathcal{W}}(w_i) \otimes \phi_{\mathcal{A}}(a_i))\|^2_{\mathcal{H}_{\mathcal{Z}}} + \lambda_1 \|C\|^2_{S_2(\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{A}}, \mathcal{H}_{\mathcal{Z}})}
$$

The minimizer $\hat{C}_{Z|W,A}$ of the loss function $\hat{\mathcal{L}}^c(C)$ is given by:

$$
\hat{C}_{Z|W,A}(\phi_{\mathcal{W}}(w) \otimes \phi_{\mathcal{A}}(a)) = \hat{\mu}_{Z|W,A}(w, a) = \sum_{i=1}^{n} \beta_i(w, a) \phi_{\mathcal{Z}}(z_i) = \Phi_{\mathcal{Z}} \boldsymbol{\beta}(w, a)
$$

where

$$
\boldsymbol{\beta}(w, a) = (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda \boldsymbol{I})^{-1} (\boldsymbol{K}_{Ww} \odot \boldsymbol{K}_{Aa})
$$

$$\Phi_{\mathcal{Z}} = \begin{bmatrix}\phi_{\mathcal{Z}}(z_1) & \dots & \phi_{\mathcal{Z}}(z_n)\end{bmatrix}$$

Second-Stage: We first consider the simplification of the population loss for two-stage regression:

$$\begin{aligned}
&\mathcal{L}^{2SR}(\varphi) = \mathbb{E}\big[\big(r(W,A) - \mathbb{E}[\varphi(Z,A)|W,A]\big)^2\big] + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}\\
&\propto \mathbb{E}\big[\mathbb{E}[\varphi(Z,A)|W,A]^2\big] - 2\int \frac{p(w)p(a)}{p(w,a)}\mathbb{E}[\varphi(Z,a)|w,a]p(w,a)dwda + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}\\
&= \mathbb{E}\big[\mathbb{E}[\varphi(Z,A)|W,A]^2\big] - 2\mathbb{E}_W\mathbb{E}_A\big[\mathbb{E}[\varphi(Z,A)|W,A]\big] + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}\\
&= \mathbb{E}\big[\langle\varphi, \mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\rangle^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}\big]\\
&- 2\mathbb{E}_W\mathbb{E}_A\big[\langle\varphi, \mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\rangle_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}\big] + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}.
\end{aligned}$$

Recall that the set $\{\tilde{z}_i, \tilde{w}_i, \tilde{a}_i\}_{i=1}^m$ denotes the second-stage data. Then, the empirical objective can be written as

$$\begin{aligned}
\hat{\mathcal{L}}^{2SR}_m(\varphi) = &\frac{1}{m}\sum_{i=1}^m \langle\varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\rangle^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}\\
&- 2\frac{1}{m(m-1)}\sum_{i=1}^m\sum_{\substack{j=1\\ j\neq i}}^m \Big\langle\varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\Big\rangle_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}} + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}
\end{aligned} \tag{10}$$

The minimizer of this objective should be in the span of the following set,

$$\varphi \in \operatorname{span}\Bigg\{\{\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\}_{i=1}^m \cup \Big\{\frac{1}{m(m-1)}\sum_{i=1}^m\sum_{\substack{j=1\\ j\neq i}}^m \hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\Big\}\Bigg\}.$$

Hence, we write

$$\varphi = \sum_{i=1}^m \alpha_i\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i) + \frac{\alpha_{m+1}}{m(m-1)}\sum_{j=1}^m\sum_{\substack{l=1\\ l\neq j}}^m \hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)$$

Let us notice the result of the following inner product which will come up a lot in our derivations,

$$\begin{aligned}
&\Big\langle\hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i), \hat{\mu}_{Z|W,A}(\tilde{w}_p,\tilde{a}_l)\otimes\phi_{\mathcal{A}}(\tilde{a}_l)\Big\rangle_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}} = \langle\hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i), \hat{\mu}_{Z|W,A}(\tilde{w}_p,\tilde{a}_l)\rangle\langle\phi_{\mathcal{A}}(\tilde{a}_i), \phi_{\mathcal{A}}(\tilde{a}_l)\rangle\\
&= \boldsymbol{\beta}(\tilde{w}_j,\tilde{a}_i)^T\Phi_{\mathcal{Z}}^T\Phi_{\mathcal{Z}}\boldsymbol{\beta}(\tilde{w}_p,\tilde{a}_l)k_{\mathcal{A}}(\tilde{a}_i,\tilde{a}_l) = \boldsymbol{\beta}(\tilde{w}_j,\tilde{a}_i)^T\boldsymbol{K}_{ZZ}\boldsymbol{\beta}(\tilde{w}_p,\tilde{a}_l)k_{\mathcal{A}}(\tilde{a}_i,\tilde{a}_l)
\end{aligned} \tag{11}$$

Next, we will calculate the terms in our second-stage regression loss individually. We begin with the squared norm of φ_0:

$$\begin{aligned}
\|\varphi\|^2 &= \langle\varphi,\varphi\rangle\\
&= \Bigg\langle\sum_{i=1}^m \alpha_i\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i) + \frac{\alpha_{m+1}}{m(m-1)}\sum_{j=1}^m\sum_{\substack{l=1\\ l\neq j}}^m \hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j),\\
&\sum_{p=1}^m \alpha_p\hat{\mu}_{Z|W,A}(\tilde{w}_p,\tilde{a}_p)\otimes\phi_{\mathcal{A}}(\tilde{a}_p) + \frac{\alpha_{m+1}}{m(m-1)}\sum_{r=1}^m\sum_{\substack{s=1\\ s\neq r}}^m \hat{\mu}_{Z|W,A}(\tilde{w}_s,\tilde{a}_r)\otimes\phi_{\mathcal{A}}(\tilde{a}_r)\Bigg\rangle\\
&= \sum_{i=1}^m\sum_{p=1}^m \alpha_i\alpha_p\langle\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i), \hat{\mu}_{Z|W,A}(\tilde{w}_p,\tilde{a}_p)\otimes\phi_{\mathcal{A}}(\tilde{a}_p)\rangle\\
&+ 2\alpha_{m+1}\frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j=1}^m\sum_{\substack{l=1\\ l\neq j}}^m \alpha_i\langle\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i), \hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\rangle
\end{aligned}$$

$$+ \alpha_{m+1}^2 \frac{1}{m^2(m-1)^2} \sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \sum_{r=1}^m \sum_{\substack{s=1 \\ s\neq r}}^m \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \hat{\mu}_{Z|W,A}(\tilde{w}_s, \tilde{a}_r) \otimes \phi_{\mathcal{A}}(\tilde{a}_r)\rangle$$

Using Equation (11), we can write

$$\begin{aligned}
&\|\varphi\|^2 = \langle \varphi, \varphi \rangle \\
&= \sum_{i=1}^m \sum_{p=1}^m \alpha_i \alpha_p \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_p, \tilde{a}_p) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_p) && (12) \\
&+ 2\alpha_{m+1} \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \alpha_i \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_j) && (13) \\
&+ \alpha_{m+1}^2 \frac{1}{m^2(m-1)^2} \sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \sum_{r=1}^m \sum_{\substack{s=1 \\ s\neq r}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_r) && (14)
\end{aligned}$$

The component in Equation (12):

$$\sum_{i=1}^m \sum_{p=1}^m \alpha_i \alpha_p \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_p, \tilde{a}_p) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_p) = \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] \alpha_{1:m}$$

where

$$\begin{aligned}
\alpha_{1:m} &= \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_m \end{bmatrix}^T \\
\boldsymbol{B} &= (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{W}} \odot \boldsymbol{K}_{A\tilde{A}})
\end{aligned}$$

The component in Equation (13):

$$\begin{aligned}
&2\alpha_{m+1} \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \alpha_i \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_j) \\
&= 2\alpha_{m+1} \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \Big(\frac{1}{m-1} \sum_{\substack{l=1 \\ l\neq j}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \Big) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_j) \\
&= 2\alpha_{m+1} \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] (\mathbf{1}/m)
\end{aligned}$$

where $\bar{\boldsymbol{B}}$ is the matrix whose j-th column is given by

$$\bar{\boldsymbol{B}}_{:,j} = \frac{1}{m-1} \sum_{\substack{l=1 \\ l\neq j}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) = \frac{1}{m-1} \sum_{\substack{l=1 \\ l\neq j}}^m (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j})$$

Finally, the third component in Equation (14):

$$\begin{aligned}
&\alpha_{m+1}^2 \frac{1}{m^2(m-1)^2} \sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \sum_{r=1}^m \sum_{\substack{s=1 \\ s\neq r}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_r) \\
&= \alpha_{m+1}^2 \frac{1}{m^2} \sum_{j=1}^m \sum_{r=1}^m \Big(\frac{1}{m-1} \sum_{\substack{l=1 \\ l\neq j}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \Big)^T \boldsymbol{K}_{ZZ} \Big(\frac{1}{m-1} \sum_{\substack{s=1 \\ s\neq r}}^m \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) \Big) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_r) \\
&= \alpha_{m+1}^2 (\mathbf{1}/m)^T \Big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) (\mathbf{1}/m)
\end{aligned}$$

As a result,

$$\|\varphi\|^2 = \langle \varphi, \varphi \rangle$$

$$
\begin{aligned}
&= \alpha_{1:m}^T \Big(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) \alpha_{1:m} + 2\alpha_{m+1} \alpha_{1:m}^T \Big(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) (\boldsymbol{1}/m) \\
&+ \alpha_{m+1}^2 (\boldsymbol{1}/m)^T \Big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) (\boldsymbol{1}/m) \\
&= \begin{bmatrix} \alpha_{1:m}^T & \alpha_{m+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}})(\boldsymbol{1}/m) \\ (\boldsymbol{1}/m)^T (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}})^T & (\boldsymbol{1}/m)^T \Big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) (\boldsymbol{1}/m) \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \qquad (15)
\end{aligned}
$$

Next, to derive the matrix-vector multiplication form for the first component of the objective given in Equation (10), consider the following:.

$$
\begin{aligned}
&\Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle \\
&= \Bigg\langle \sum_{l=1}^m \alpha_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) + \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Bigg\rangle \\
&= \sum_{l=1}^m \alpha_l \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l), \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle \\
&+ \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle \\
&= \sum_{l=1}^m \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i) + \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_i) \\
&= \sum_{l=1}^m \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i) + \frac{\alpha_{m+1}}{m} \sum_{j=1}^m \Big(\frac{1}{m-1} \sum_{\substack{l=1 \\ l \neq j}}^m \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \Big)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_i) \\
&= \Big[(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) \alpha_{1:m} \Big]_i + \alpha_{m+1} \Big[(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) (\boldsymbol{1}/m) \Big]_i \\
&= \Bigg[\begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) (\boldsymbol{1}/m) \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \Bigg]_i
\end{aligned}
$$

As a result, the first component in Equation (10) is given by

$$
\begin{aligned}
&\frac{1}{m} \sum_{i=1}^m \Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle^2 \\
&= \frac{1}{m} \begin{bmatrix} \alpha_{1:m}^T & \alpha_{m+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\boldsymbol{1}}{m})^T \big(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \big)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & \big(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \big) \frac{\boldsymbol{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \quad (16)
\end{aligned}
$$

Lastly, for the second component in Equation (10), we note that

$$
\begin{aligned}
&\frac{1}{m(m-1)} \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \sum_{l=1}^m \alpha_l \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l), \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&+ \frac{\alpha_{m+1}}{m^2(m-1)^2} \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \sum_{r=1}^m \sum_{\substack{s=1 \\ s \neq r}}^m \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_s, \tilde{a}_r) \otimes \phi_{\mathcal{A}}(\tilde{a}_r), \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{\substack{j=1 \\ j \neq i}}^m \sum_{l=1}^m \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\alpha_{m+1}}{m^2(m-1)^2} \sum_{i=1}^{m} \sum_{\substack{j=1\\ j\neq i}}^{m} \sum_{r=1}^{m} \sum_{\substack{s=1\\ s\neq r}}^{m} \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_r, \tilde{a}_i) \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \Big(\frac{1}{m-1} \sum_{\substack{j=1\\ j\neq i}}^{m} \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) \Big) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i) \\
&+ \frac{\alpha_{m+1}}{m^2} \sum_{i=1}^{m} \sum_{r=1}^{m} \Big(\frac{1}{m-1} \sum_{\substack{s=1\\ s\neq r}}^{m} \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) \Big)^T \boldsymbol{K}_{ZZ} \Big(\frac{1}{m-1} \sum_{\substack{j=1\\ j\neq i}}^{m} \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) \Big) k_{\mathcal{A}}(\tilde{a}_r, \tilde{a}_i) \\
&= \frac{1}{m} \alpha_{1:m}^T \Big(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) \mathbf{1} + \alpha_{m+1} \frac{1}{m^2} \mathbf{1}^T \Big(\bar{\boldsymbol{B}} \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) \mathbf{1} \\
&= \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T (\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) \frac{\mathbf{1}}{m} \end{bmatrix}
\end{aligned} \tag{17}
$$

Using Equation (15), (16) and (17), we can write

$$
\begin{aligned}
&\hat{\mathcal{L}}_m^{2SR}(\varphi) = \frac{1}{m} \sum_{i=1}^{m} \langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \rangle^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&- 2 \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{\substack{j=1\\ j\neq i}}^{m} \Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} + \lambda_2 \|\varphi_0\|^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{m} \begin{bmatrix} \alpha_{1:m}^T & \alpha_{m+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\mathbf{1}}{m})^T \big(\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \big)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) \frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \\
&- 2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T \Big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) \frac{\mathbf{1}}{m} \end{bmatrix} \\
&+ \lambda_2 \begin{bmatrix} \alpha_{1:m}^T & \alpha_{m+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}) \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T (\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}})^T & (\frac{\mathbf{1}}{m})^T \Big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big) \frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}
\end{aligned} \tag{18}
$$

The optimal coefficients $\{\alpha_{1:m}, \alpha_{m+1}\}$ can be found by setting the derivative of Equation (18) to zero. With these optimal coefficients, let $\hat{\varphi}_{\lambda_2,m}$ denote the minimizer of $\hat{\mathcal{L}}_m^{2SR}(\varphi)$. Using $\hat{\varphi}_{\lambda_2,m}$, we can estimate $\mathbb{E}[Y \hat{\varphi}_{\lambda_2,m}(Z,a)|A=a]$, thus obtaining the desired ATE function estimation. First, observe that

$$
\begin{aligned}
\mathbb{E}[Y \hat{\varphi}_{\lambda_2,m}(Z,a)|A=a] &= \mathbb{E}[Y \langle \hat{\varphi}_{\lambda_2,m}, \phi_{\mathcal{Z}}(Z) \otimes \phi_{\mathcal{A}}(a) \rangle | A=a] \\
&= \Big\langle \hat{\varphi}_{\lambda_2,m}, \mathbb{E}[Y \phi_{\mathcal{Z}}(Z)|A=a] \otimes \phi_{\mathcal{A}}(a) \Big\rangle \\
&= \langle \hat{\varphi}_{\lambda_2,m}, C_{YZ|A} \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a) \rangle
\end{aligned}
$$

where $C_{YZ|A}$ is the conditional mean operator, i.e., $C_{YZ|A}\phi_{\mathcal{A}}(a) = \mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A=a]$ and it is estimated by kernel ridge regression:

$$
\begin{aligned}
\hat{C}_{YZ|A} &= \arg\min_{C} \frac{1}{n} \sum_{i=1}^{n} \|y_i \phi_{\mathcal{Z}}(z_i) - C_{YZ|A} \phi_{\mathcal{A}}(a_i)\|^2 + \lambda_3 \|C_{YZ|A}\|^2 \\
&= \arg\min_{C} \frac{1}{n} \|\Phi_{\mathcal{Z}} \text{diag}(\boldsymbol{Y}) - C_{YZ|A} \Phi_{\mathcal{A}}\|^2 + \lambda_3 \|C_{YZ|A}\|^2.
\end{aligned} \tag{19}
$$

The solution for Equation (19) is given by

$$
\hat{C}_{YZ|A} \phi_{\mathcal{A}}(a) = \Phi_{\mathcal{Z}} \text{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa}
$$

As a result,

$$
\mathbb{E}[Y \hat{\varphi}_{\lambda_2,m}(Z,a)|A=a] = \langle \hat{\varphi}_{\lambda_2,m}, \hat{C}_{YZ|A} \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a) \rangle
$$

$$\begin{aligned}
&= \Big\langle \sum_{l=1}^{m} \alpha_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l), \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a) \Big\rangle \\
&+ \Big\langle \alpha_{m+1} \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a) \Big\rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l), \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a) \rangle \\
&+ \alpha_{m+1} \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a) \rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l), \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) \rangle \langle \phi_{\mathcal{A}}(\tilde{a}_l), \phi_{\mathcal{A}}(a) \rangle \\
&+ \alpha_{m+1} \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j), \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) \rangle \langle \phi_{\mathcal{A}}(\tilde{a}_j), \phi_{\mathcal{A}}(a) \rangle \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \Phi_{\mathcal{Z}}^T \Phi_{\mathcal{Z}} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} k_{\mathcal{A}}(\tilde{a}_l, a) \\
&+ \alpha_{m+1} \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \Phi_{\mathcal{Z}}^T \Phi_{\mathcal{Z}} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} k_{\mathcal{A}}(\tilde{a}_l, a) \\
&+ \alpha_{m+1} \frac{1}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \boldsymbol{K}_{ZZ} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \alpha_{1:m}^T \Big(\boldsymbol{B}^T \big(\boldsymbol{K}_{ZZ} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} \big) \odot \boldsymbol{K}_{\tilde{A}a} \Big) \\
&+ \alpha_{m+1} \Big(\bar{\boldsymbol{B}}^T \big(\boldsymbol{K}_{ZZ} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} \big) \odot \boldsymbol{K}_{\tilde{A}a} \Big) \frac{\boldsymbol{1}}{m}
\end{aligned}$$

As a result,

$$\begin{aligned}
f_{\mathrm{ATE}}(a) = \alpha_{1:m}^T &\Big(\boldsymbol{B}^T \big(\boldsymbol{K}_{ZZ} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} \big) \odot \boldsymbol{K}_{\tilde{A}a} \Big) \\
&+ \alpha_{m+1} \Big(\bar{\boldsymbol{B}}^T \big(\boldsymbol{K}_{ZZ} \mathrm{diag}(\boldsymbol{Y}) [\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I}]^{-1} \boldsymbol{K}_{Aa} \big) \odot \boldsymbol{K}_{\tilde{A}a} \Big) \frac{\boldsymbol{1}}{m}
\end{aligned}$$

□

One observation that is a result of the Algorithm (4.1) is that we can compute the bridge function in the closed-form as stated in the following remark.

Remark 11.1. *Given the optimal coefficients $\{\alpha_{1:m}, \alpha_{m+1}\}$ from Algorithm (4.1), the bridge function can be written in closed-form as*

$$\hat{\varphi}_{\lambda_2,m}(z,a) = \alpha_{1:m}^T [(\boldsymbol{B}^T \boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}] + \alpha_{m+1} \Big(\frac{\boldsymbol{1}}{m}\Big)^T [(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}]$$

Proof.

$$\hat{\varphi}_{\lambda_2,m}(z,a) = \langle \hat{\varphi}_{\lambda_2,m}, \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \rangle$$

$$
\begin{aligned}
&= \Bigg\langle \sum_{l=1}^{m} \alpha_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) + \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \Bigg\rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l), \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \rangle \\
&+ \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l), \phi_{\mathcal{Z}}(z) \rangle k_{\mathcal{A}}(\tilde{a}_l, a) + \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j), \phi_{\mathcal{Z}}(z) \rangle k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \sum_{l=1}^{m} \alpha_l \langle \Phi_{\mathcal{Z}} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l), \phi_{\mathcal{Z}}(z) \rangle k_{\mathcal{A}}(\tilde{a}_l, a) + \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \langle \Phi_{\mathcal{Z}} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j), \phi_{\mathcal{Z}}(z) \rangle k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{K}_{Zz}^T \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l) k_{\mathcal{A}}(\tilde{a}_l, a) + \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \boldsymbol{K}_{Zz}^T \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_l}) k_{\mathcal{A}}(\tilde{a}_l, a) \\
&+ \frac{\alpha_{m+1}}{m^2} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j}) k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_l}) k_{\mathcal{A}}(\tilde{a}_l, a) \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} \Big(\frac{1}{m-1} \sum_{\substack{l=1 \\ l \neq j}}^{m} \boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j} \Big) k_{\mathcal{A}}(\tilde{a}_j, a) \\
&= \alpha_{1:m}^T [(\boldsymbol{B}^T \boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}] + \alpha_{m+1} \Big(\frac{\boldsymbol{1}}{m} \Big)^T [(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}]
\end{aligned}
$$

□

11.2 Conditional Dose-Response Curve Algorithm

In this section, we provide the derivation of Algorithm (4.2). For the completeness, we write the algorithm first.

Algorithm (Algorithm (4.2))**.** *Let $\{z_i, w_i, a_i\}_{i=1}^{n}$ and $\{\tilde{w}_i, \tilde{a}_i\}_{i=1}^{m}$ be the first-stage and second-stage data, respectively, and $(\lambda_1, \lambda_2, \lambda_3, \zeta)$ be the regularization parameters. Furthermore, let $\{\alpha_i\}_{i=1}^{m+1}$ be the minimizer of the following cost function for a given a':*

$$
\begin{aligned}
&\hat{\mathcal{L}}_m^{2SR}(\boldsymbol{\alpha}) = \\
&\frac{1}{m} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\boldsymbol{1}}{m})^T \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\boldsymbol{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} k_{\mathcal{A}}(a', a')^2 \\
&- 2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\boldsymbol{1}}{m} \\ (\frac{\boldsymbol{1}}{m})^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \frac{\boldsymbol{1}}{m} \end{bmatrix} k_{\mathcal{A}}(a', a') \\
&+ \lambda_2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\boldsymbol{1}}{m} \\ (\frac{\boldsymbol{1}}{m})^T [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]^T & (\frac{\boldsymbol{1}}{m})^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \frac{\boldsymbol{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} k_{\mathcal{A}}(a', a')
\end{aligned}
$$

where $\tilde{\boldsymbol{B}}$ is the matrix whose j-th column is given by $\tilde{\boldsymbol{B}}_{:,j} = \sum_{\substack{l=1 \\ l \neq j}}^{m} (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\theta_l \boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j})$ and $\theta_l = [(\boldsymbol{K}_{\tilde{A}\tilde{A}} + m\zeta \boldsymbol{I})^{-1} \boldsymbol{K}_{\tilde{A}a'}]_l$. Furthermore, the kernel matrices and the matrix $\boldsymbol{B}$ are as defined in Algorithm

(4.1). Then, the conditional dose-response curve estimation can be written in the closed-form as

$$\begin{aligned}f_{ATT}(a,a') &= \alpha_{1:m}^T\Big(\boldsymbol{B}^T\big(\boldsymbol{K}_{ZZ}\,diag(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}\big)\odot\boldsymbol{K}_{\tilde{A}a}\Big)k_{\mathcal{A}}(a',a')\\ &+\alpha_{m+1}\Big(\tilde{\boldsymbol{B}}^T\big(\boldsymbol{K}_{ZZ}\,diag(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}\big)\odot\boldsymbol{K}_{\tilde{A}a}\Big)\frac{\boldsymbol{1}}{m}k_{\mathcal{A}}(a',a')\end{aligned}$$

Derivation of Algorithm (4.2). Let $r(W,a,a')$ denote $\frac{p(W,a')p(a)}{p(W,a)p(a')}$. We aim to find the optimum of the following loss function:

$$\mathcal{L}^{2\mathrm{SR}}(\varphi)=\mathbb{E}[(r(W,A,a')-\mathbb{E}[\varphi(Z,A,a')|W,A])^2]+\lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{A}}}$$

in which we assume that the bridge function φ is the RKHS $\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{A}}$. Similar to the dose-response algorithm, this loss function cannot be directly optimized as it involves the conditional expectation $\mathbb{E}[\varphi(Z,A,a')|W,A]$. Following (Xu and Gretton, 2024), we similarly employ a two-stage regression approach. The first-stage is identical to the one used for the dose-response curve as described in S.M. (11.1). In this stage, we find the conditional mean embedding:

$$\hat{\mu}_{Z|W,A}(w,a)=\sum_{i=1}^n\beta_i(w,a)\phi_{\mathcal{Z}}(z_i)=\Phi_{\mathcal{Z}}\boldsymbol{\beta}(w,a)$$

where

$$\begin{aligned}\boldsymbol{\beta}(w,a)&=(\boldsymbol{K}_{WW}\odot\boldsymbol{K}_{AA}+n\lambda_1\boldsymbol{I})^{-1}(\boldsymbol{K}_{Ww}\odot\boldsymbol{K}_{Aa})\\ \Phi_{\mathcal{Z}}&=\begin{bmatrix}\phi_{\mathcal{Z}}(z_1) & \dots & \phi_{\mathcal{Z}}(z_n)\end{bmatrix}\end{aligned}$$

For the second-stage, we note the following:

$$\begin{aligned}&\mathbb{E}[(r(W,A,a')-\mathbb{E}[\varphi(Z,A,a')|W,A])^2]\\ &=\mathbb{E}[\mathbb{E}[\varphi(Z,A,a')|W,A]^2]-2\mathbb{E}[r(W,A,a')\mathbb{E}[\varphi(Z,A,a')|W,A]]+\text{const.}\\ &=\mathbb{E}[\mathbb{E}[\varphi(Z,A,a')|A,W]^2]-2\int\frac{p(w,a')p(a)}{p(w,a)p(a')}\mathbb{E}[\varphi(Z,A,a')|A=a,W=w]p(w,a)dwda+\text{const.}\\ &=\mathbb{E}[\mathbb{E}[\varphi(Z,A,a')|W=w,A=a]^2]-2\int p(w|a')p(a)\mathbb{E}[\varphi(Z,A,a')|W=w,A=a]dwda+\text{const.}\\ &=\mathbb{E}[\mathbb{E}[\varphi(Z,A,a')|W=w,A=a]^2]-2\mathbb{E}_{W|A'=a'}\mathbb{E}_A[\mathbb{E}[\varphi(Z,A,a')|W=w,A=a]]+\text{const.}\end{aligned}$$

Recall that under Assumption 5.1-(1), $\mu_{Z|W,A}(w,a)=C_{Z|W,A}(\phi_{\mathcal{W}}(w)\otimes\phi_{\mathcal{A}}(a))$. Hence, we can write,

$$\begin{aligned}&\mathbb{E}[(r(W,A,a')-\mathbb{E}[\varphi(Z,A,a')|W,A])^2]\\ &=\mathbb{E}[\langle\varphi,\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\otimes\phi_{\mathcal{A}}(a')\rangle^2]\\ &-2\mathbb{E}_{W|A'=a'}\mathbb{E}_A[\langle\varphi,C_{Z|W,A}(\phi_{\mathcal{W}}(W)\otimes\phi_{\mathcal{A}}(A))\otimes\phi_{\mathcal{A}}(A)\otimes\phi_{\mathcal{A}}(a')\rangle]+\text{const.}\\ &=\mathbb{E}[\langle\varphi,\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\otimes\phi_{\mathcal{A}}(a')\rangle^2]\\ &-2\mathbb{E}_A[\langle\varphi,C_{Z|W,A}(\mathbb{E}_{W|A=a'}[\phi_{\mathcal{W}}(W)]\otimes\phi_{\mathcal{A}}(A))\otimes\phi_{\mathcal{A}}(A)\otimes\phi_{\mathcal{A}}(a')\rangle]+\text{const.}\end{aligned}$$

This expectation can be estimated from the second stage data $\{\tilde{w}_i,\tilde{z}_i,\tilde{a}_i\}$ using the following expression, up to a constant factor:

$$\begin{aligned}&\mathbb{E}[(r(W,A,a')-\mathbb{E}[\varphi(Z,A,a')|W,A])^2]\\ &\approx\frac{1}{m}\sum_{i=1}^m\langle\varphi,\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\rangle^2\\ &-\frac{2}{m}\sum_{j=1}^m\langle\varphi,\hat{C}_{Z|W,A}(\hat{\mathbb{E}}_{W|A=a'}[\phi_{\mathcal{W}}(W)]\otimes\phi_{\mathcal{A}}(\tilde{a}_j))\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\rangle.\end{aligned}$$

Here, $\hat{\mathbb{E}}_{W|A'=a'}[\phi_{\mathcal{W}}(W)]$ is the estimation of conditional mean embedding for $p(W|A=a')$, and can be expressed in the closed-form as the result of kernel ridge regression with regularization parameter ζ:

$$\hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] = \Phi_{\mathcal{W}}(\boldsymbol{K}_{\tilde{A}\tilde{A}} + m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{A}a'} = \sum_{i=1}^{m}\theta_i\phi_{\mathcal{W}}(\tilde{w}_i) = \Phi_{\mathcal{W}}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = (\boldsymbol{K}_{\tilde{A}\tilde{A}} + m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{A}a'}$ and $\Phi_{\mathcal{W}} = \begin{bmatrix}\phi_{\mathcal{W}}(\tilde{w}_1) & \dots & \phi_{\mathcal{W}}(\tilde{w}_m)\end{bmatrix}$. Hence, the least-squares loss can be estimated by second-stage data $\{\tilde{w}_i, \tilde{a}_i\}_{i=1}^m$ as follows:

$$\begin{aligned}
&\mathbb{E}[(r(W,A,a') - \mathbb{E}[\varphi(Z,A,a')|W,A])^2] \\
&\approx \frac{1}{m}\sum_{i=1}^{m}\langle\varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\rangle^2 \\
&- 2\frac{1}{m}\sum_{j=1}^{m}\langle\varphi, \hat{C}_{Z|W,A}(\sum_{\substack{i=1\\ i\neq j}}^{m}\theta_i\phi_{\mathcal{W}}(\tilde{w}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_j))\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\rangle \\
&= \frac{1}{m}\sum_{i=1}^{m}\langle\varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\rangle^2 \\
&- 2\frac{1}{m}\sum_{j=1}^{m}\sum_{\substack{i=1\\ i\neq j}}^{m}\langle\varphi, \hat{C}_{Z|W,A}(\theta_i\phi_{\mathcal{W}}(\tilde{w}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_j))\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\rangle \\
&= \frac{1}{m}\sum_{i=1}^{m}\langle\varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\rangle^2 \\
&- 2\frac{1}{m}\sum_{j=1}^{m}\sum_{\substack{i=1\\ i\neq j}}^{m}\langle\varphi, \theta_i\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\rangle.
\end{aligned}$$

As a result, we can write the sample loss for two-stage regression with Tikhonov regularization as

$$\begin{aligned}
\hat{\mathcal{L}}_m^{\mathrm{2SR}}(\varphi) &= \frac{1}{m}\sum_{i=1}^{m}\langle\varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\rangle^2 \\
&- 2\frac{1}{m}\sum_{j=1}^{m}\sum_{\substack{i=1\\ i\neq j}}^{m}\langle\varphi, \theta_i\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\rangle + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{A}}}. \qquad (20)
\end{aligned}$$

We can see that the minimizer of this objective should be in the span of the following set,

$$\varphi \in \mathrm{span}\left\{\{\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\}_{i=1}^m \cup \left\{\frac{1}{m}\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\theta_j\hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a')\right\}\right\}.$$

Hence, we write

$$\varphi = \sum_{i=1}^{m}\alpha_i\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m}\sum_{j=1}^{m}\sum_{\substack{l=1\\ l\neq j}}^{m}\theta_l\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')$$

Let us notice the result of the following inner product which will come up a lot

$$\begin{aligned}
&\Big\langle\hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i)\otimes\phi_{\mathcal{A}}(a'), \hat{\mu}_{Z|W,A}(\tilde{w}_p,\tilde{a}_l)\otimes\phi_{\mathcal{A}}(\tilde{a}_l)\otimes\phi_{\mathcal{A}}(a')\Big\rangle_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{A}}} \\
&= \langle\hat{\mu}_{Z|W,A}(\tilde{w}_j,\tilde{a}_i), \hat{\mu}_{Z|W,A}(\tilde{w}_p,\tilde{a}_l)\rangle\langle\phi_{\mathcal{A}}(\tilde{a}_i),\phi_{\mathcal{A}}(\tilde{a}_l)\rangle\langle\phi_{\mathcal{A}}(a'),\phi_{\mathcal{A}}(a')\rangle \\
&= \boldsymbol{\beta}(\tilde{w}_j,\tilde{a}_i)^T\Phi_{\mathcal{Z}}^T\Phi_{\mathcal{Z}}\boldsymbol{\beta}(\tilde{w}_p,\tilde{a}_l)k_{\mathcal{A}}(\tilde{a}_i,\tilde{a}_l)k_{\mathcal{A}}(a',a') \\
&= \boldsymbol{\beta}(\tilde{w}_j,\tilde{a}_i)^T\boldsymbol{K}_{ZZ}\boldsymbol{\beta}(\tilde{w}_p,\tilde{a}_l)k_{\mathcal{A}}(\tilde{a}_i,\tilde{a}_l)k_{\mathcal{A}}(a',a') \qquad (21)
\end{aligned}$$

We will now compute the individual terms in $\mathcal{L}_m^{\text{2SR}}(\varphi)$ one by one. We start by the squared norm of φ:

$$\begin{aligned}
&\|\varphi\|^2 = \langle \varphi, \varphi \rangle \\
&= \Bigg\langle \sum_{i=1}^m \alpha_i \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m} \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \\
&\sum_{p=1}^m \alpha_p \hat{\mu}_{Z|W,A}(\tilde{w}_p, \tilde{a}_p) \otimes \phi_{\mathcal{A}}(\tilde{a}_p) \otimes \phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m} \sum_{r=1}^m \sum_{\substack{s=1 \\ s \neq r}}^m \theta_s \hat{\mu}_{Z|W,A}(\tilde{w}_s, \tilde{a}_r) \otimes \phi_{\mathcal{A}}(\tilde{a}_r) \otimes \phi_{\mathcal{A}}(a') \Bigg\rangle \\
&= \sum_{i=1}^m \sum_{p=1}^m \alpha_i \alpha_p \langle \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a'), \hat{\mu}_{Z|W,A}(\tilde{w}_p, \tilde{a}_p) \otimes \phi_{\mathcal{A}}(\tilde{a}_p) \otimes \phi_{\mathcal{A}}(a') \rangle \\
&+ 2\frac{\alpha_{m+1}}{m} \sum_{i=1}^m \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \alpha_i \langle \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a'), \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a') \rangle \\
&+ \frac{\alpha_{m+1}^2}{m^2} \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \sum_{r=1}^m \sum_{\substack{s=1 \\ s \neq r}}^m \langle \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \theta_s \hat{\mu}_{Z|W,A}(\tilde{w}_s, \tilde{a}_r) \otimes \phi_{\mathcal{A}}(\tilde{a}_r) \otimes \phi_{\mathcal{A}}(a') \rangle
\end{aligned}$$

Using Equation (21), we can write

$$\begin{aligned}
&\|\varphi\|^2 = \langle \varphi, \varphi \rangle \\
&= \sum_{i=1}^m \sum_{p=1}^m \alpha_i \alpha_p \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_p, \tilde{a}_p) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_p) k_{\mathcal{A}}(a', a') && (22) \\
&+ 2\frac{\alpha_{m+1}}{m} \sum_{i=1}^m \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \alpha_i \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_j) k_{\mathcal{A}}(a', a') && (23) \\
&+ \frac{\alpha_{m+1}^2}{m^2} \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \sum_{r=1}^m \sum_{\substack{s=1 \\ s \neq r}}^m \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \boldsymbol{K}_{ZZ} \theta_s \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_r) k_{\mathcal{A}}(a', a') && (24)
\end{aligned}$$

The component in Equation (22) is equal to:

$$\sum_{i=1}^m \sum_{p=1}^m \alpha_i \alpha_p \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_p, \tilde{a}_p) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_p) k_{\mathcal{A}}(a', a') = \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \alpha_{1:m} k_{\mathcal{A}}(a', a')$$

where

$$\begin{aligned}
\alpha_{1:m} &= \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_m \end{bmatrix}^T \\
\boldsymbol{B} &= (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{W}} \odot \boldsymbol{K}_{A\tilde{A}})
\end{aligned}$$

The component in Equation (23) is equal to:

$$\begin{aligned}
&2\frac{\alpha_{m+1}}{m} \sum_{i=1}^m \sum_{j=1}^m \sum_{\substack{l=1 \\ l \neq j}}^m \alpha_i \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_j) k_{\mathcal{A}}(a', a') \\
&= 2\frac{\alpha_{m+1}}{m} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i)^T \boldsymbol{K}_{ZZ} \Big(\sum_{\substack{l=1 \\ l \neq j}}^m \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \Big) k_{\mathcal{A}}(\tilde{a}_i, \tilde{a}_j) k_{\mathcal{A}}(a', a') \\
&= 2\alpha_{m+1} \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] (\boldsymbol{1}/m) k_{\mathcal{A}}(a', a')
\end{aligned}$$

where

$$\tilde{\boldsymbol{B}}_{:,j} = \sum_{\substack{l=1 \\ l \neq j}}^m \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) = \sum_{l=1}^m (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda \boldsymbol{I})^{-1} (\theta_l \boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j})$$

$$= (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} \Big(\sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j} \Big)$$

Finally, the third component in Equation (24):

$$\begin{aligned}
&\frac{\alpha_{m+1}^2}{m^2} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \sum_{r=1}^{m} \sum_{\substack{s=1 \\ s \neq r}}^{m} \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)^T \boldsymbol{K}_{ZZ} \theta_s \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_r) k_{\mathcal{A}}(a', a') \\
&= \frac{\alpha_{m+1}^2}{m^2} \sum_{j=1}^{m} \sum_{r=1}^{m} \Big(\sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \Big)^T \boldsymbol{K}_{ZZ} \Big(\sum_{\substack{s=1 \\ s \neq r}}^{m} \theta_s \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) \Big) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_r) k_{\mathcal{A}}(a', a') \\
&= \alpha_{m+1}^2 (\mathbf{1}/m)^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] (\mathbf{1}/m) k_{\mathcal{A}}(a', a')
\end{aligned}$$

As a result,

$$\begin{aligned}
&\|\varphi\|^2 = \langle \varphi, \varphi \rangle \\
&= \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] \alpha_{1:m} k_{\mathcal{A}}(a', a') + 2\alpha_{m+1} \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] (\mathbf{1}/m) k_{\mathcal{A}}(a', a') \\
&+ \alpha_{m+1}^2 (\mathbf{1}/m)^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] (\mathbf{1}/m) k_{\mathcal{A}}(a', a') \\
&= \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}](\mathbf{1}/m) \\ (\mathbf{1}/m)^T [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]^T & (\mathbf{1}/m)^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \Big] (\mathbf{1}/m) \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} k_{\mathcal{A}}(a', a')
\end{aligned} \tag{25}$$

Next, to derive the matrix-vector multiplication form for the first component of the objective given in Equation (20), consider the following:

$$\begin{aligned}
&\Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle \\
&= \Bigg\langle \sum_{l=1}^{m} \alpha_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \\
&\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Bigg\rangle \\
&= \sum_{l=1}^{m} \alpha_l \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a'), \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \Big\langle \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \Big(\sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \Big)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(\tilde{w}_i, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_j, \tilde{a}_i) k_{\mathcal{A}}(a', a') \\
&= \Big[[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \alpha_{1:m} \Big]_i k_{\mathcal{A}}(a', a') + \alpha_{m+1} \Big[[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] (\mathbf{1}/m) \Big]_i k_{\mathcal{A}}(a', a') \\
&= \left[\begin{matrix} [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}](\mathbf{1}/m)] \end{matrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \right]_i k_{\mathcal{A}}(a', a')
\end{aligned}$$

As a result, the first component in Equation (20) is given by

$$\frac{1}{m} \sum_{i=1}^{m} \Big\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle^2$$

$$= \frac{1}{m} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\mathbf{1}}{m})^T \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T \end{bmatrix} \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \quad \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big] \frac{\mathbf{1}}{m}\big] \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} k_{\mathcal{A}}(a', a')^2 \tag{26}$$

Lastly, for the second component in Equation (20), we note that

$$\begin{aligned}
&\frac{1}{m} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \Big\langle \varphi, \theta_j \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{l=1}^{m} \alpha_l \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a'), \theta_j \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle \\
&+ \frac{\alpha_{m+1}}{m^2} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{r=1}^{m} \sum_{\substack{s=1 \\ s \neq r}}^{m} \Big\langle \theta_s \hat{\mu}_{Z|W,A}(\tilde{w}_s, \tilde{a}_r) \otimes \phi_{\mathcal{A}}(\tilde{a}_r) \otimes \phi_{\mathcal{A}}(a'), \theta_j \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \theta_j \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m^2} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sum_{r=1}^{m} \sum_{\substack{s=1 \\ s \neq r}}^{m} \theta_s \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r)^T \boldsymbol{K}_{ZZ} \theta_j \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) k_{\mathcal{A}}(\tilde{a}_r, \tilde{a}_i) k_{\mathcal{A}}(a', a') \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \Big(\sum_{\substack{j=1 \\ j \neq i}}^{m} \theta_j \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) \Big) k_{\mathcal{A}}(\tilde{a}_l, \tilde{a}_i) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m^2} \sum_{i=1}^{m} \sum_{r=1}^{m} \Big(\sum_{\substack{s=1 \\ s \neq r}}^{m} \theta_s \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) \Big)^T \boldsymbol{K}_{ZZ} \Big(\sum_{\substack{j=1 \\ j \neq i}}^{m} \theta_j \boldsymbol{\beta}(\tilde{w}_j, \tilde{a}_i) \Big) k_{\mathcal{A}}(\tilde{a}_r, \tilde{a}_i) k_{\mathcal{A}}(a', a') \\
&= \frac{1}{m} \alpha_{1:m}^T \Big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \mathbf{1} k_{\mathcal{A}}(a', a') + \alpha_{m+1} \frac{1}{m^2} \mathbf{1}^T \Big[\tilde{\boldsymbol{B}} \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \mathbf{1} k_{\mathcal{A}}(a', a') \\
&= \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \frac{\mathbf{1}}{m} \end{bmatrix} k_{\mathcal{A}}(a', a')
\end{aligned} \tag{27}$$

Now, we are ready to combine our findings and write the loss function in terms of matrix-vector multiplications. Using Equation (25), (26) and (27), the loss function can be expressed as

$$\begin{aligned}
&\hat{\mathcal{L}}_m^{2SR}(\varphi) = \frac{1}{m} \sum_{i=1}^{m} \langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \rangle^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&- 2 \frac{1}{m} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \Big\langle \varphi, \theta_j \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{A}}} + \lambda_2 \|\varphi_0\|^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{m} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\mathbf{1}}{m})^T \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T \end{bmatrix} \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \quad \big[\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big] \frac{\mathbf{1}}{m}\big] \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} k_{\mathcal{A}}(a', a')^2 \\
&- 2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \frac{\mathbf{1}}{m} \end{bmatrix} k_{\mathcal{A}}(a', a') \\
&+ \lambda_2 \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}] \frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T [\boldsymbol{B}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]^T & (\frac{\mathbf{1}}{m})^T \Big[\tilde{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big] \frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} k_{\mathcal{A}}(a', a')
\end{aligned} \tag{28}$$

The optimal coefficients $\{\alpha_{1:m}, \alpha_{m+1}\}$ can be found by setting the derivative of Equation (28) to zero. With these optimal coefficients, let $\hat{\varphi}_{\lambda_2,m}$ denote the minimizer of $\hat{\mathcal{L}}_m^{2SR}(\varphi)$. Using $\hat{\varphi}_{\lambda_2,m}$, we can estimate

$\mathbb{E}[Y\hat{\varphi}_{\lambda_2,m}(Z,a,a')|A=a]$. First, observe that

$$\begin{aligned}\mathbb{E}[Y\hat{\varphi}_{\lambda_2,m}(Z,a,a')|A=a] &= \mathbb{E}[Y\langle\hat{\varphi}_{\lambda_2,m},\phi_{\mathcal{Z}}(Z)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\rangle|A=a]\\ &= \Big\langle\hat{\varphi}_{\lambda_2,m},\mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A=a]\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\Big\rangle\\ &\approx \langle\hat{\varphi}_{\lambda_2,m},\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\rangle\end{aligned}$$

where $\hat{C}_{YZ|A}$ is the estimation of the conditional mean $\mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A=\cdot]$, as used in the dose-response curve algorithm. It is found by kernel ridge regression:

$$\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a) = \Phi_{\mathcal{Z}}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}$$

Thus,

$$\begin{aligned}\hat{\mathbb{E}}[Y\hat{\varphi}_{\lambda_2,m}(Z,a,a')|A=a] &= \langle\hat{\varphi}_{\lambda_2,m},\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\rangle\\ &= \Big\langle\sum_{l=1}^{m}\alpha_l\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_l)\otimes\phi_{\mathcal{A}}(\tilde{a}_l)\otimes\phi_{\mathcal{A}}(a'),\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\Big\rangle\\ &+\Big\langle\frac{\alpha_{m+1}}{m}\sum_{j=1}^{m}\sum_{\substack{l=1\\l\neq j}}^{m}\theta_l\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a'),\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\Big\rangle\\ &= \sum_{l=1}^{m}\alpha_l\langle\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_l)\otimes\phi_{\mathcal{A}}(\tilde{a}_l)\otimes\phi_{\mathcal{A}}(a'),\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\rangle\\ &+\frac{\alpha_{m+1}}{m}\sum_{j=1}^{m}\sum_{\substack{l=1\\l\neq j}}^{m}\theta_l\langle\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a'),\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\rangle\\ &= \sum_{l=1}^{m}\alpha_l\langle\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_l),\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\rangle\langle\phi_{\mathcal{A}}(\tilde{a}_l),\phi_{\mathcal{A}}(a)\rangle\langle\phi_{\mathcal{A}}(a'),\phi_{\mathcal{A}}(a')\rangle\\ &+\frac{\alpha_{m+1}}{m}\sum_{j=1}^{m}\sum_{\substack{l=1\\l\neq j}}^{m}\langle\theta_l\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j),\hat{C}_{YZ|A}\phi_{\mathcal{A}}(a)\rangle\langle\phi_{\mathcal{A}}(\tilde{a}_j),\phi_{\mathcal{A}}(a)\rangle\langle\phi_{\mathcal{A}}(a'),\phi_{\mathcal{A}}(a')\rangle\\ &= \sum_{l=1}^{m}\alpha_l\boldsymbol{\beta}(\tilde{w}_l,\tilde{a}_l)^T\Phi_{\mathcal{Z}}^T\Phi_{\mathcal{Z}}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}k_{\mathcal{A}}(\tilde{a}_l,a)k_{\mathcal{A}}(a',a')\\ &+\frac{\alpha_{m+1}}{m}\sum_{j=1}^{m}\sum_{\substack{l=1\\l\neq j}}^{m}\theta_l\boldsymbol{\beta}(\tilde{w}_l,\tilde{a}_j)^T\Phi_{\mathcal{Z}}^T\Phi_{\mathcal{Z}}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}k_{\mathcal{A}}(\tilde{a}_j,a)k_{\mathcal{A}}(a',a')\\ &= \sum_{l=1}^{m}\alpha_l\boldsymbol{\beta}(\tilde{w}_l,\tilde{a}_l)^T\boldsymbol{K}_{ZZ}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}k_{\mathcal{A}}(\tilde{a}_l,a)k_{\mathcal{A}}(a',a')\\ &+\frac{\alpha_{m+1}}{m}\sum_{j=1}^{m}\sum_{\substack{l=1\\l\neq j}}^{m}\theta_l\boldsymbol{\beta}(\tilde{w}_l,\tilde{a}_j)^T\boldsymbol{K}_{ZZ}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}k_{\mathcal{A}}(\tilde{a}_j,a)k_{\mathcal{A}}(a',a')\\ &= \alpha_{1:m}^T\Big(\boldsymbol{B}^T\big(\boldsymbol{K}_{ZZ}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}\big)\odot\boldsymbol{K}_{\tilde{A}a}\Big)k_{\mathcal{A}}(a',a')\\ &+\alpha_{m+1}\Big(\tilde{\boldsymbol{B}}^T\big(\boldsymbol{K}_{ZZ}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_3\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}\big)\odot\boldsymbol{K}_{\tilde{A}a}\Big)\frac{\mathbf{1}}{m}k_{\mathcal{A}}(a',a')\end{aligned}$$

The conditional dose-response curve can therefore be expressed in the closed-form as

$$\begin{aligned}f_{\mathrm{ATT}}(a,a') &= \alpha_{1:m}^T\Big(\boldsymbol{B}^T\big(\boldsymbol{K}_{ZZ}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_2\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}\big)\odot\boldsymbol{K}_{\tilde{A}a}\Big)k_{\mathcal{A}}(a',a')\\ &+\alpha_{m+1}\Big(\tilde{\boldsymbol{B}}^T\big(\boldsymbol{K}_{ZZ}\mathrm{diag}(\boldsymbol{Y})[\boldsymbol{K}_{AA}+n\lambda_2\boldsymbol{I}]^{-1}\boldsymbol{K}_{Aa}\big)\odot\boldsymbol{K}_{\tilde{A}a}\Big)\frac{\mathbf{1}}{m}k_{\mathcal{A}}(a',a')\end{aligned}$$

□

Remark 11.2. *Given the optimal coefficients $\{\alpha_{1:m}, \alpha_{m+1}\}$ from Algorithm (4.2), the bridge function can be written in the closed-form as*

$$\hat{\varphi}_{\lambda_2,m}(z,a,a') = k_{\mathcal{A}}(a',a')\alpha_{1:m}^T[(\boldsymbol{B}^T\boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}] + k_{\mathcal{A}}(a',a')\alpha_{m+1}\Big(\frac{\mathbf{1}}{m}\Big)^T[(\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}]$$

Proof.

$$\begin{aligned}
\hat{\varphi}_{\lambda_2,m}(z,a,a') &= \langle \hat{\varphi}_{\lambda_2,m}, \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle \\
&= \Bigg\langle \sum_{l=1}^{m} \alpha_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \\
&\phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a') \Bigg\rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a'), \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \phi_{\mathcal{Z}}(z) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l), \phi_{\mathcal{Z}}(z)\rangle k_{\mathcal{A}}(\tilde{a}_l, a) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j), \phi_{\mathcal{Z}}(z)\rangle k_{\mathcal{A}}(\tilde{a}_j, a) k_{\mathcal{A}}(a', a') \\
&= \sum_{l=1}^{m} \alpha_l \langle \Phi_{\mathcal{Z}} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l), \phi_{\mathcal{Z}}(z)\rangle k_{\mathcal{A}}(\tilde{a}_l, a) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \langle \Phi_{\mathcal{Z}} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j), \phi_{\mathcal{Z}}(z)\rangle k_{\mathcal{A}}(\tilde{a}_j, a) k_{\mathcal{A}}(a', a') \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{K}_{Zz}^T \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l) k_{\mathcal{A}}(\tilde{a}_l, a) k_{\mathcal{A}}(a', a') + \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \boldsymbol{K}_{Zz}^T \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) k_{\mathcal{A}}(\tilde{a}_j, a) k_{\mathcal{A}}(a', a') \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_l}) k_{\mathcal{A}}(\tilde{a}_l, a) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j}) k_{\mathcal{A}}(\tilde{a}_j, a) k_{\mathcal{A}}(a', a') \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} (\boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_l}) k_{\mathcal{A}}(\tilde{a}_l, a) k_{\mathcal{A}}(a', a') \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \boldsymbol{K}_{Zz}^T (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I})^{-1} \Big(\sum_{\substack{l=1 \\ l \neq j}}^{m} \theta_l \boldsymbol{K}_{W\tilde{w}_l} \odot \boldsymbol{K}_{A\tilde{a}_j} \Big) k_{\mathcal{A}}(\tilde{a}_j, a) k_{\mathcal{A}}(a', a') \\
&= k_{\mathcal{A}}(a',a')\alpha_{1:m}^T[(\boldsymbol{B}^T\boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}] + k_{\mathcal{A}}(a',a')\alpha_{m+1}\Big(\frac{\mathbf{1}}{m}\Big)^T[(\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{Zz}) \odot \boldsymbol{K}_{\tilde{A}a}]
\end{aligned}$$

□

12 CONSISTENCY RESULTS

In this section, we provide the consistency result of our proposed method. We make the following assumptions on the kernels and on the noise between the outcome and the treatment.

Assumption 12.1 (Replication of Assumption (5.2))**.** *For $\mathcal{F} \in \{\mathcal{A}, \mathcal{W}, \mathcal{Z}\}$, we assume that*

- *$\mathcal{F}$ is a Polish space;*
- *$k_{\mathcal{F}}(f, .)$, is continuous for almost every $f \in \mathcal{F}$;*
- *$k_{\mathcal{F}}(f, .)$, is bounded by κ for almost every $f \in \mathcal{F}$, i.e.,*

$$\sup_{f \in \mathcal{F}} \|k_{\mathcal{F}}(f, .)\|_{\mathcal{H}_{\mathcal{F}}} \leq \kappa.$$

- *There exists $R, \sigma > 0$ such that for all $q \geq 2$, P_A-almost surely,*

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid A])^q \mid A] \leq \frac{1}{2} q! \sigma^2 R^{q-2}. \tag{29}$$

The last assumption is a Bernstein moment condition used to control the noise of the observations (see Caponnetto and De Vito (2007); Fischer and Steinwart (2020) for more details). If Y is almost surely bounded, then the condition is automatically satisfied. It is possible to prove that the Bernstein condition is equivalent to sub-exponentiality, see Mollenhauer et al. (2022, Remark 4.9).

12.1 Consistency Results for Dose-Response Curve

12.1.1 Assumptions for ATE

Under Assumption (5.1-2), there exists a solution to the bridge equation within the RKHS $\mathcal{H}_{\mathcal{ZA}}$. However, this solution might not be unique. We therefore introduce the following minimum norm solution in the set of valid bridge functions.

Definition 12.2 (Bridge solution with minimum RKHS norm)**.** *We define*

$$\bar{\varphi}_0 = \underset{\varphi \in \mathcal{H}_{\mathcal{ZA}}}{\arg\min} \|\varphi\|_{\mathcal{H}_{\mathcal{ZA}}} \qquad s.t. \qquad \mathbb{E}[\varphi(Z, A)|W, A] = r(W, A),$$

with $r(W, A) = \frac{p(W)p(A)}{p(W,A)}$.

Under Assumption (5.1-2), $\bar{\varphi}_0$ is well-defined. We will show that the estimator from stage 2 converges to $\bar{\varphi}_0$ in RKHS norm and we will therefore be able to obtain consistency guarantees for the dose-response function.

Remark 12.3 (Uniqueness of the bridge function)**.** *We notice that previous works on Proximal Causal Learning (Mastouri et al., 2021; Xu et al., 2021; Singh, 2023; Cui et al., 2024; Wu et al., 2024) require the bridge solution to be unique. However, we show that this assumption is not needed and convergence to the minimum norm bridge solution is enough to obtain consistency on the estimation of the dose-response curve. Our results build upon recent advances in instrumental variable regression with kernel methods Meunier et al. (2024).*

We now introduce the covariance operators associated to stages 1, 2 and 3.

Definition 12.4. *The covariance operators are defined as*

1. *(Stage 1)* $\Sigma_1 := \mathbb{E}\left[\phi_{\mathcal{WA}}(W, A) \otimes \phi_{\mathcal{WA}}(W, A)\right], \qquad \phi_{\mathcal{WA}}(W, A) = \phi_{\mathcal{W}}(W) \otimes \phi_{\mathcal{A}}(A);$
2. *(Stage 2)* $\Sigma_2 := \mathbb{E}_{W,A}\left[\left(\left(\mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A)\right) \otimes \left(\mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A)\right)\right)\right];$
3. *(Stage 3)* $\Sigma_3 := \mathbb{E}[\phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(A)].$

$\Sigma_1, \Sigma_2, \Sigma_3$ are self-adjoint and positive semi-definite operators. Under Assumption (12.1), they are trace class, and therefore compact, which implies that they have a countable spectrum (Steinwart and Christmann, 2008).

The next proposition relates $\bar{\varphi}_0$ to Σ_2.

Proposition 12.5. *Under Assumption (5.1-2), $\bar{\varphi}_0$ is well-defined and is the unique element of $\mathcal{H}_{\mathcal{ZA}}$ satisfying $\mathbb{E}[\bar{\varphi}_0(Z,A)|W,A] = r(W,A)$ and such that $\bar{\varphi}_0 \in \operatorname{null}(\Sigma_2)^\perp$.*

Proof. Note that the bridge equation, for an element $\varphi \in \mathcal{H}_{\mathcal{ZA}}$, can be written as

$$r(W,A) = E[\varphi(Z,A)|W,A] = \langle \varphi, \mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)\rangle_{\mathcal{H}_{\mathcal{ZA}}} = A\varphi,$$

by using the reproducing property and introducing the operator $A : \mathcal{H}_{\mathcal{ZA}} \to \mathcal{L}_2(\mathcal{W} \times \mathcal{A}, p_{W,A}), \varphi \mapsto \langle \varphi, \mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)\rangle_{\mathcal{H}_{\mathcal{ZA}}}$ where $\mathcal{L}_2(\mathcal{W} \times \mathcal{A}, p_{W,A})$ denotes the square integrable functions with respect to the measure $p(W,A)$. By Assumption (5.1-2), $A^{-1}(\{r\}) \subseteq \mathcal{H}_{\mathcal{ZA}}$ is not empty. Fix φ an element of $A^{-1}(\{r\})$. Since $\mathcal{H}_{\mathcal{ZA}} = \operatorname{null}(A) \oplus \operatorname{null}(A)^\perp$ there exists a unique pair $(\varphi', \varphi'') \in \operatorname{null}(A)^\perp \times \operatorname{null}(A)$ such that $\varphi = \varphi' + \varphi''$. Since $\varphi \in A^{-1}(\{r\})$ and $\varphi'' \in \operatorname{null}(A)$, we have:

$$r = A\varphi = A\varphi' + A\varphi'' = A\varphi'.$$

Therefore $\varphi' \in A^{-1}(\{r\})$. Furthermore, $\|\varphi\|^2_{\mathcal{H}_{\mathcal{ZA}}} = \|\varphi'\|^2_{\mathcal{H}_{\mathcal{ZA}}} + \|\varphi''\|^2_{\mathcal{H}_{\mathcal{ZA}}} \geq \|\varphi'\|^2_{\mathcal{H}_{\mathcal{ZA}}}$. This proves that the minimum norm solution in $\mathcal{H}_{\mathcal{ZA}}$ exists and is uniquely defined as φ' and belongs to $\operatorname{null}(A)^\perp \cap A^{-1}(\{r\})$. We then show that $\operatorname{null}(A)^\perp \cap A^{-1}(\{r\})$ contains only one element. Assume that there exists $\varphi, \tilde{\varphi} \in \operatorname{null}(A)^\perp \cap A^{-1}(\{r\})$, then $A(\varphi - \tilde{\varphi}) = r - r = 0$, therefore $\varphi - \tilde{\varphi} \in \operatorname{null}(A)$. But since we also have $\varphi - \tilde{\varphi} \in \operatorname{null}(A)^\perp$, it implies $\varphi = \tilde{\varphi}$. To conclude, observe that A is such that $\Sigma_2 = A^*A$, therefore $\operatorname{null}(A) = \operatorname{null}(A^*A) = \operatorname{null}(\Sigma_2)$. □

To characterize the smoothness of the target functions for each respective stage we employ the following source assumption.

Assumption 12.6 (Replication of Assumption (5.3))**.** *We assume that the following conditions hold:*

1. *There exists a constant $B_1 < \infty$ such that for a given $\beta_1 \in (1,3]$,*

$$\|C_{Z|W,A}\Sigma_1^{-\frac{\beta_1-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{WA}},\mathcal{H}_{\mathcal{Z}})} \leq B_1$$

2. *There exists a constant $B_2 < \infty$ such that for a given $\beta_2 \in (1,3]$,*

$$\|\Sigma_2^{-\frac{\beta_2-1}{2}}\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \leq B_2.$$

3. *There exists a constant $B_3 < \infty$ such that for a given $\beta_3 \in (1,3]$,*

$$\|C_{YZ|A}\Sigma_3^{-\frac{\beta_3-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})} \leq B_3.$$

This assumption is referred to as the source condition in the literature (Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020). It measures the smoothness of the regression functions with respect to the covariance operators. The inverse covariance operators have to be understood as Moore–Penrose pseudoinverses (Ben-Israel and Greville, 2006). In particular, $\Sigma_2^{\frac{\beta_2-1}{2}}\Sigma_2^{-\frac{\beta_2-1}{2}} = P_2$, with P_2 the orthogonal projection onto $\operatorname{null}\left(\Sigma_2^{\frac{\beta_2-1}{2}}\right)^\perp = \operatorname{null}(\Sigma_2)^\perp$. Combined with Proposition (12.5), we obtain the following result.

Proposition 12.7. *Under Assumption (5.1-2), $\bar{\varphi}_0 = \Sigma_2^{\frac{\beta_2-1}{2}}\Sigma_2^{-\frac{\beta_2-1}{2}}\bar{\varphi}_0$.*

Remark 12.8 (Smoothness and minimum norm solution)**.** *$\bar{\varphi}_0$ is the unique RKHS solution to the bridge equation such that Proposition (12.7) holds. Indeed by Proposition (12.5), $\bar{\varphi}_0$ is the unique RKHS solution to the bridge equation such that $\bar{\varphi}_0 \in \operatorname{null}(\Sigma_2)^\perp$ and therefore such that $\bar{\varphi}_0 = P_2\bar{\varphi}_0$. This crucial property will allow us to show that our bridge estimator converges to $\bar{\varphi}_0$.*

The next assumption characterize the effective dimension of the RKHSs associated to each stage. It is a standard assumption on the eigenvalue decay of the covariance operators (see more details in Caponnetto and De Vito (2007); Fischer and Steinwart (2020)).

Assumption 12.9. *We assume the following conditions hold*

1. *Let $(\lambda_{1,i})_{i \geq 1}$ be the eigenvalues of Σ_1. For some constant $c_1 > 0$ and parameter $p_1 \in (0,1]$ and for all $i \geq 1$,*
$$\lambda_{1,i} \leq c_1 i^{-1/p_1}.$$
2. *Let $(\lambda_{2,i})_{i \geq 1}$ be the eigenvalues of Σ_2. For some constant $c_2 > 0$ and parameter $p_2 \in (0,1]$ and for all $i \geq 1$,*
$$\lambda_{2,i} \leq c_2 i^{-1/p_2}.$$
3. *Let $(\lambda_{3,i})_{i \geq 1}$ be the eigenvalues of Σ_3. For some constant $c_3 > 0$ and parameter $p_3 \in (0,1]$ and for all $i \geq 1$,*
$$\lambda_{3,i} \leq c_3 i^{-1/p_3}.$$

12.1.2 Proof sketch for ATE

We provide non-asymptotic uniform consistency guarantees for the dose-response curve. Below we provide a proof sketch.

First Stage regression. The estimator from stage 1 aims at estimating the conditional mean embedding $\mu_{Z|W,A} = \mathbb{E}[\phi_{\mathcal{Z}}(Z) \mid W, A]$. We recall that under the well-specifiedness assumption (Assumption (5.1-1)), we have $\mu_{Z|W,A}(\cdot,\cdot) = C_{Z|W,A}(\phi_{\mathcal{W}}(\cdot) \otimes \phi_{\mathcal{A}}(\cdot))$, with $C_{Z|W,A} \in S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})$. We point out that Assumption (5.1-1) is equivalent to the assumption that $\mu_{Z|W,A}(\cdot,\cdot)$ belong to the vector-valued RKHS associated to the vector-valued kernel $K((w,a),(w',a')) = \langle \phi_{\mathcal{WA}}(w,a), \phi_{\mathcal{WA}}(w',a') \rangle_{\mathcal{H}_{\mathcal{WA}}} \operatorname{Id}_{\mathcal{H}_{\mathcal{Z}}}$, where $\operatorname{Id}_{\mathcal{H}_{\mathcal{Z}}}$ denotes the identity operator in $\mathcal{H}_{\mathcal{Z}}$ (see Li et al. (2022) for a detailed discussion).

Given the regularization parameter $\lambda_1 > 0$, we recall that the objective to learn the conditional mean embedding operator is,

$$\hat{\mathcal{L}}^c(C) = \frac{1}{n}\sum_{i=1}^n \|\phi_{\mathcal{Z}}(z_i) - C(\phi_{\mathcal{W}}(w_i) \otimes \phi_{\mathcal{A}}(a_i))\|^2_{\mathcal{H}_{\mathcal{Z}}} + \lambda_1 \|C\|^2_{S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})}, \qquad C \in S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}}),$$

whose minimizer is denoted as,

$$\hat{C}_{Z|W,A} = \underset{C \in S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})}{\arg\min}\, \hat{\mathcal{L}}^c(C).$$

The conditional mean embedding is then approximated as

$$\hat{\mu}_{Z|W,A}(w,a) = \hat{C}_{Z|W,A}(\phi_{\mathcal{W}}(w) \otimes \phi_{\mathcal{A}}(a)), \qquad w \in \mathcal{W}, \quad a \in \mathcal{A}.$$

We bound $\|\hat{C}_{Z|W,A} - C_{Z|W,A}\|_{S_2}$ in S.M. (Sec. 12.1.3), using the main result from (Li et al., 2022). We then convert this bound to a bound on $\|\hat{\mu}_{Z|W,A} - \mu_{Z|W,A}\|_\infty$, that will allow us to obtain the uniform consistency of the dose-response function.

Second Stage regression. We recall that for the second stage we have the following loss at the population level:

$$\mathcal{L}^{2SR}(\varphi) = \mathbb{E}\big[(r(W,A) - \mathbb{E}[\varphi(Z,A) \mid W,A])^2\big], \qquad \varphi \in \mathcal{H}_{\mathcal{ZA}}.$$

We showed in Eq. (3) that $\mathcal{L}^{2SR}$ can be equivalently written as

$$\mathcal{L}^{2SR}(\varphi) = \mathbb{E}\big[\mathbb{E}[\varphi(Z,A) \mid W,A]^2\big] - 2\mathbb{E}_W\mathbb{E}_A\big[\mathbb{E}[\varphi(Z,A) \mid W,A]\big] + \text{const}.$$

We introduce the regularized version of the population loss, for $\lambda_2 > 0$,

$$\mathcal{L}^{2SR}_{\lambda_2}(\varphi) = \mathcal{L}^{2SR}(\varphi) + \lambda_2 \|\varphi\|^2_{\mathcal{H}_{\mathcal{ZA}}}.$$

Let us introduce $g_2 = \mathbb{E}_W\mathbb{E}_A\left[\mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)\right]$.

Proposition 12.10. *g_2 can be alternatively written as*

$$g_2 = \mathbb{E}[r(W,A)\mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)].$$

Furthermore, for any element $\varphi_0 \in \mathcal{H}_{\mathcal{ZA}}$ solution to the bridge equation, we have $g_2 = \Sigma_2\varphi_0$. Finally,

$$\varphi_{\lambda_2} := \underset{\varphi \in \mathcal{H}_{\mathcal{ZA}}}{\arg\min}\, \mathcal{L}^{2SR}_{\lambda_2}(\varphi) = \left(\Sigma_2 + \lambda_2 Id_{\mathcal{H}_{\mathcal{ZA}}}\right)^{-1} g_2.$$

Proof. The first part follows from the same derivations as in Eq. (3). For the second part, if φ_0 is such that $r(W,A) = \langle \varphi_0, \mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}$, then,

$$g_2 = \mathbb{E}_{W,A}[r(W,A)\mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)] = \Sigma_2 \varphi_0.$$

For the final part, notice that for $\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}}$, by the reproducing property,

$$\begin{aligned}\mathcal{L}^{2SR}_{\lambda_2}(\varphi) &= \mathbb{E}_{W,A}\left[\langle \varphi, \mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)\rangle^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\right] - 2\mathbb{E}_W\mathbb{E}_A\left[\langle \varphi, \mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\right] + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \\ &\quad + \text{const} \\ &= \langle \varphi, \Sigma_2 \varphi\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} - 2\langle \varphi, g_2\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \text{const} \\ &= \langle \varphi, (\Sigma_2 + \lambda_2 \operatorname{Id})\varphi\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} - 2\langle \varphi, g_2\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \text{const}.\end{aligned}$$

We conclude by setting the Fréchet derivative to 0. □

We would now like to consider the empirical version of the previous loss. The standard kernel ridge regression estimator would obtain an estimator by replacing both the population covariance Σ_2 and the term g_2 by their empirical counterpart in Proposition (12.10). However, as g_2 takes a different form, we need to specify how we build its empirical counterpart.

- Option 1: take the empirical counterpart of $g_2 = \Sigma_2\varphi_0$; this is not feasible as it would require the knowledge of a bridge function.
- Option 2: take the empirical counterpart of $g_2 = \mathbb{E}[r(W,A)\mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A)]$; this would require the estimation of the density ratio $r(W,A)$ and would be inefficient in high dimension.
- Option 3 (**ours**): take the empirical counterpart of $g_2 = \mathbb{E}_W\mathbb{E}_A\left[\mu_{Z|W,A} \otimes \phi_{\mathcal{A}}(A)\right]$; this is the estimator suggested in Section (4.2) that allows us to by-pass density ratio estimation.

We therefore introduce

$$\begin{aligned}\bar{\Sigma}_{2,m} &= \frac{1}{m}\sum_{i=1}^m \Big(\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\Big) \otimes \Big(\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\Big), \\ \bar{g}_{2,m} &= \frac{1}{m(m-1)}\sum_{\substack{i,j \\ i\neq j}}^m \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i).\end{aligned}$$

However, as we do not directly observe the conditional mean embedding $\mu_{Z|W,A}$, we plug-in its approximation obtained in the first stage regression. This leads us to,

$$\begin{aligned}\hat{\Sigma}_{2,m} &= \frac{1}{m}\sum_{i=1}^m \Big(\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\Big) \otimes \Big(\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\Big), \\ \hat{g}_{2,m} &= \frac{1}{m(m-1)}\sum_{\substack{i,j \\ i\neq j}}^m \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i).\end{aligned}$$

Let us then introduce the following empirical losses,

$$\bar{\mathcal{L}}^{2SR}_m(\varphi) = \frac{1}{m}\sum_{i=1}^m \langle \varphi, \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\rangle^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} - \frac{2}{m(m-1)}\sum_{\substack{i,j=1 \\ j\neq i}}^m \langle \varphi, \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}},$$

$$\hat{\mathcal{L}}^{2SR}_m(\varphi) = \frac{1}{m}\sum_{i=1}^m \langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, , \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\rangle^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} - \frac{2}{m(m-1)}\sum_{\substack{i,j=1 \\ j\neq i}}^m \langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)\rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \lambda_2\|\varphi_0\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}.$$

We can observe that the minimizers of the objective functions are given by

$$\varphi_{\lambda_2} = (\Sigma_2 + \lambda_2 I)^{-1} g_2 = \operatorname*{arg\,min}_{\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}}} \mathcal{L}^{2SR}_{\lambda_2}(\varphi),$$

$$\bar{\varphi}_{\lambda_2,m} = (\bar{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{2,m} = \operatorname*{arg\,min}_{\varphi \in \mathcal{H}_{\mathcal{ZA}}} \bar{\mathcal{L}}_m^{2SR}(\varphi),$$
$$\hat{\varphi}_{\lambda_2,m} = (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\hat{g}_{2,m} = \operatorname*{arg\,min}_{\varphi \in \mathcal{H}_{\mathcal{ZA}}} \hat{\mathcal{L}}_m^{2SR}(\varphi).$$

$\hat{\varphi}_{\lambda_2,m}$ is the final estimator presented in Section (4.2). In S.M. (Sec. 12.1.4), we show the convergence of $\hat{\varphi}_{\lambda_2,m}$ to the minimum norm bridge $\bar{\varphi}_0$ introduced in Definition (12.2). $\bar{\varphi}_{\lambda_2,m}$ and φ_{λ_2} are introduced for theoretical reasons. Indeed, we will consider the following decomposition,

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \le \|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZA}}} + \|\bar{\varphi}_{\lambda_2,m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZA}}} + \|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}.$$

Third Stage regression. The estimator from stage 3 aims at estimating $\Psi(A) := \mathbb{E}[Y\phi_{\mathcal{Z}}(Z)|A]$. We recall that under the well-specifiedness assumption (Assumption (5.1-3)), we have $\Psi(\cdot) = C_{YZ|A}\phi_{\mathcal{A}}(\cdot)$, with $C_{YZ|A} \in S_2(\mathcal{H}_{\mathcal{A}}, \mathcal{H}_{\mathcal{Z}})$.

Given the regularization parameter $\lambda_3 > 0$, we recall that the objective to learn $C_{YZ|A}$ is,

$$\hat{\mathcal{L}}_3(C) = \frac{1}{t}\sum_{i=1}^{t} \|\bar{y}_i\phi_{\mathcal{Z}}(\bar{z}_i) - C\phi_{\mathcal{A}}(\bar{a}_i))\|_{\mathcal{H}_{\mathcal{Z}}}^2 + \lambda_3\|C\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})}^2, \qquad C \in S_2(\mathcal{H}_{\mathcal{A}}, \mathcal{H}_{\mathcal{Z}}),$$

where $\{(\bar{a}_i, \bar{z}_i, \bar{y}_i)\}_{i=1}^t$ can re-use data from stage 1 or 2 or both. The minimizer is denoted as,

$$\hat{C}_{YZ|A} = \operatorname*{arg\,min}_{C \in S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})} \hat{\mathcal{L}}_3(C),$$

leading to

$$\hat{\Psi}(a) = \hat{C}_{YZ|A}\phi_{\mathcal{A}}(a), \qquad a \in \mathcal{A}.$$

We note that $\Psi(A)$ can be interpreted as the conditional kernel mean embedding of the random variable (Y, Z) given A with the following kernel: $k_{YZ}((y, z), (y', z')) = yy'k_{\mathcal{Z}}(z, z')$, $(y, z), (y', z') \in \mathbb{R} \times \mathcal{Z}$. Indeed, the canonical feature map of k_{YZ} is $y\phi_{\mathcal{Z}}(z)$ for $(y, z) \in \mathbb{R} \times \mathcal{Z}$. We could therefore proceed as for stage 1 and apply results from (Li et al., 2022). However, the analysis of (Li et al., 2022) would require k_{YZ} to be bounded on the support of (Y, Z) and therefore requires Y to be almost surely bounded. Instead, in Section (12.1.5), we apply results from (Li et al., 2024) which generalize consistency guarantees for conditional mean operator learning to general vector-valued regression. The results applies with the weaker assumption that Y is sub-exponential (Assumption (12.1), Equation (29)).

Uniform consistency guarantees for ATE. We recall that after obtaining the estimators $\hat{\varphi}_{\lambda_2,m}$ from stage 2 and $\hat{\Psi}$ from stage 3, we have

$$\hat{f}_{ATE}(\cdot) = \langle \hat{\varphi}_{\lambda_2,m}, \hat{\Psi}(\cdot) \otimes \phi_{\mathcal{A}}(\cdot)\rangle_{\mathcal{H}_{\mathcal{ZA}}}.$$

On the other hand, under Assumption (5.1),

$$f_{ATE}(\cdot) = \langle \bar{\varphi}_0, \Psi(\cdot) \otimes \phi_{\mathcal{A}}(\cdot)\rangle_{\mathcal{H}_{\mathcal{ZA}}}.$$

For any $a \in \mathcal{A}$, we apply the following decomposition,

$$\begin{aligned}
&|\hat{f}_{ATE}(a) - f_{ATE}(a)| = |\langle \hat{\varphi}_{\lambda_2,m}, \hat{\Psi}(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}} - \langle \bar{\varphi}_0, \Psi(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}}| \\
&\ = |\langle \hat{\varphi}_{\lambda_2,m}, (\hat{\Psi} - \Psi)(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}} + \langle (\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0), \Psi(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}}| \\
&\ = |\langle \hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0, (\hat{\Psi} - \Psi)(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}} + \langle \bar{\varphi}_0, (\hat{\Psi} - \Psi)(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}} + \langle (\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0), \Psi(a) \otimes \phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}}| \\
&\ \le \kappa\left(\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} + \|\bar{\varphi}_0\|\|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{ZA}}} + \|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\|\Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}}\right) \qquad (30)
\end{aligned}$$

By plugging the consistency results from stage 2 and 3, we obtain the final bound that leads to Theorem (5.4). See S.M. (Sec. 12.1.6) for details. In the next sections, we detail each step of the proof.

12.1.3 First-Stage Regression Consistency Result

We adapt Li et al. (2022, Theorem 2) to our setting.

Theorem 12.11 (Theorem 2 Li et al. (2022))**.** *Suppose Assumptions (5.1-1), (12.1), (12.6-1) and (12.9-1) hold and take $\lambda_1 = \Theta\left(n^{-\frac{1}{\beta_1+p_1}}\right)$. There is a constant $J_1 > 0$ independent of $n \geq 1$ and $\delta \in (0,1)$ such that*

$$\left\|\hat{C}_{Z|W,A} - C_{Z|W,A}\right\|_{S_2(\mathcal{H}_{\mathcal{WA}},\mathcal{H}_{\mathcal{Z}})} \leq J_1 \log(4/\delta)\left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}} =: r_1(\delta, n, \beta_1, p_1),$$

is satisfied for sufficiently large $n \geq 1$ with probability at least $1-\delta$.

Proof. We apply Li et al. (2022, Theorem 2-case 2.), with $\gamma = 1$ which corresponds to the Hilbert-Schmidt norm $S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})$. As we focus on the well-specified setting with $\beta_1 \geq 1$, we can apply case 2. of Li et al. (2022, Theorem 2) as in their paper $\alpha \leq 1$, hence $\beta_1 + p_1 \geq \alpha$. Note that Li et al. (2022, Theorem 2) applies under the assumption that $k_{\mathcal{Z}}$ is bounded, which is the case under Assumption 12.1. We note that in Li et al. (2022, Theorem 2), the bound is valid for $\beta_1 \in (1,2]$ while we allow for $\beta_1 \in (1,3]$ (see Remark 12.14 below). □

Corollary 12.12. *Under the same assumptions as Theorem 12.11, with $\lambda_1 = \Theta\left(n^{-\frac{1}{\beta_1+p_1}}\right)$, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\sup_{(w,a)\in\mathcal{W}\times\mathcal{A}} \|\hat{\mu}_{Z|W,A}(w,a) - \mu_{Z|W,A}(w,a)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \kappa^2 r_1(\delta, n, \beta_1, p_1).$$

Proof. For any $(w,a) \in \mathcal{W}\times\mathcal{A}$, under Assumption (12.1), we have

$$\|\phi_{\mathcal{W}}(w)\otimes\phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{WA}}} = \|\phi_{\mathcal{W}}(w)\|_{\mathcal{H}_{\mathcal{W}}}\|\phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{A}}} \leq \kappa^2.$$

As a result, we observe that,

$$\begin{aligned}\|\hat{\mu}_{Z|W,A}(w,a) - \mu_{Z|W,A}(w,a)\|_{\mathcal{H}_{\mathcal{Z}}} &= \|(\hat{C}_{Z|W,A} - C_{Z|W,A})\phi_{\mathcal{W}}(w)\otimes\phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{Z}}} \\ &\leq \|\hat{C}_{Z|W,A} - C_{Z|W,A}\|_{S_2(\mathcal{H}_{\mathcal{WA}},\mathcal{H}_{\mathcal{Z}})}\|\phi_{\mathcal{W}}(w)\otimes\phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{WA}}} \\ &\leq \kappa^2\|\hat{C}_{Z|W,A} - C_{Z|W,A}\|_{S_2(\mathcal{H}_{\mathcal{WA}},\mathcal{H}_{\mathcal{Z}})},\end{aligned}$$

and the conclusion follows from Theorem (12.11). □

12.1.4 Second-Stage Regression Consistency Results

We recall that we consider the following decomposition,

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \leq \|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZA}}} + \|\bar{\varphi}_{\lambda_2,m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZA}}} + \|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}.$$

We first consider an upper bound for $\|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}$.

Lemma 12.13. *Suppose that Assumption (12.6-2.) holds with parameter $\beta_2 \in (1,3]$. Then, for any $\lambda_2 > 0$,*

$$\|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \leq B_2\lambda_2^{\frac{\beta_2-1}{2}}.$$

Proof. We saw in Proposition (12.10) that

$$\varphi_{\lambda_2} = (\Sigma_2 + \lambda_2\,\mathrm{Id})^{-1} g_2 = (\Sigma_2 + \lambda_2\,\mathrm{Id})^{-1}\Sigma_2\bar{\varphi}_0 = \bar{\varphi}_0 - \lambda_2(\Sigma_2 + \lambda_2\,\mathrm{Id})^{-1}\bar{\varphi}_0.$$

Therefore, under Assumption (12.6-2.),

$$\|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} = \lambda_2\left\|(\Sigma_2 + \lambda_2\,\mathrm{Id})^{-1}\bar{\varphi}_0\right\|_{\mathcal{H}_{\mathcal{ZA}}} \leq B_2\lambda_2\left\|(\Sigma_2 + \lambda_2\,\mathrm{Id})^{-1}\Sigma_2^{\frac{\beta_2-1}{2}}\right\|_{op},$$

where we used Proposition (12.7) and $\|.\|_{op}$ denotes the *operator norm.*. Note that, by Lemma (12.46),

$$\left\| (\Sigma_2 + \lambda_2 \operatorname{Id})^{-1} \Sigma_2^{\frac{\beta_2-1}{2}} \right\|_{op} = \sup_{i\geq 1} \frac{\lambda_{2,i}^{\frac{\beta_2-1}{2}}}{\lambda_{2,i}+\lambda_2} \leq \lambda_2^{\frac{\beta_2-1}{2}-1},$$

as long as $\frac{\beta_2-1}{2} \in (0,1]$, i.e. $\beta_2 \in (1,3]$. By merging the bounds, we obtain the final result. □

Remark 12.14. *The commonly known saturation effect of Tikhonov regularization comes for the approximation error bound. As demonstrated above, the range of smoothness is limited to $\beta_2 \leq 3$. However, past works on kernel ridge regression (e.g. Fischer and Steinwart (2020)) or kernel PCL (e.g. Mastouri et al. (2021); Singh (2023)) observed a saturation effect at $\beta_2 = 2$. It was observed in Meunier et al. (2023, Remark 7 & Proposition 7) – see also Blanchard and Mücke (2018) – that saturation happens at $\beta_2 = 2$ when we measure the error in the L_2-norm while saturation happens at $\beta_2 = 3$ when we measure the error in the RKHS norm, as seen in the previous proof. As the error is measured in RKHS norm in both works Mastouri et al. (2021); Singh (2023), they can apply the same reasoning to extend their results from the range $\beta_2 \in (1,2]$ to the range $\beta_2 \in (1,3]$.*

We will need the following result to pursue our proof.

Lemma 12.15. *For any $\lambda_2 > 0$, $\|\varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}$.*

Proof. We saw in Proposition 12.10 that

$$\|\varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = \left\| (\Sigma_2 + \lambda_2 \operatorname{Id})^{-1} \Sigma_2 \bar{\varphi}_0 \right\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq \left\| (\Sigma_2 + \lambda_2 \operatorname{Id})^{-1} \Sigma_2 \right\|_{op} \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}.$$

□

Our proof of the convergence result relies on an Hoeffding concentration inequality (Corollary 12.48) and a Bernstein concentration inequality (Theorem 12.49) for Hilbert space-valued random variables. We introduce the effective dimension for the stage 2 error: for $\lambda_2 > 0$, $\mathcal{N}(\lambda_2) := \operatorname{Tr}((\Sigma_2 + \lambda_2 \operatorname{Id})^{-1}\Sigma_2)$ Caponnetto and De Vito (2007).

Proposition 12.16 (Lemma 11 & Lemma 13 Fischer and Steinwart (2020))**.** *Under Assumption (12.9-2), there is a constant $D > 0$ such that the following inequality is satisfied, for $\lambda_2 > 0$, $\mathcal{N}(\lambda_2) \leq D\lambda_2^{-p_2}$. Furthermore, we have the equality,*

$$\mathbb{E}\left[\left\| (\Sigma_2 + \lambda_2 \operatorname{Id})^{-1/2} \mu_{Z|W,A}(W,A) \otimes \phi_{\mathcal{A}}(A) \right\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}^2 \right] = \mathcal{N}(\lambda_2).$$

Lemma 12.17. *Let us introduce $g_{\lambda_2} = \log\left(2e\mathcal{N}(\lambda_2)\frac{\|\Sigma_2\|_{op}+\lambda_2}{\|\Sigma_2\|_{op}}\right)$. Suppose Assumption (12.1) holds. Then, with probability at least $1-\delta$ for all $\delta \in (0,1)$, for $m \geq 8\kappa^4 \log(2/\delta) g_{\lambda_2} \lambda_2^{-1}$,*

$$\|\bar{\varphi}_{\lambda_2,m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq \frac{3}{\sqrt{\lambda_2}} \left(\log(2/\delta) \sqrt{\frac{32}{m}\left(\mathcal{N}(\lambda_2)\kappa^4 \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}^2 + \frac{\kappa^8 \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}^2}{m\lambda_2} \right)} + \frac{\|g_2 - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}}{\sqrt{\lambda_2}} \right).$$

Proof. We decompose the error as

$$\begin{aligned}
\|\bar{\varphi}_{\lambda_2,m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} &= \left\| (\bar{\Sigma}_{2,m} + \lambda_2 I)^{-1} \big(\bar{g}_{2,m} - (\bar{\Sigma}_{2,m} + \lambda_2 I)\varphi_{\lambda_2} \big) \right\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \\
&\leq \left\| (\Sigma_2 + \lambda_2 I)^{-1/2} \right\|_{op} \left\| (\Sigma_2 + \lambda_2 I)^{1/2} (\bar{\Sigma}_{2,m} + \lambda_2 I)^{-1} (\Sigma_2 + \lambda_2 I)^{1/2} \right\|_{op} \\
&\times \left\| (\Sigma_2 + \lambda_2 I)^{-1/2} \big(\bar{g}_{2,m} - \bar{\Sigma}_{2,m}\varphi_{\lambda_2} - \underbrace{\lambda_2 \varphi_{\lambda_2}}_{=g_2 - \Sigma_2 \varphi_{\lambda_2}} \big) \right\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \\
&\leq \lambda_2^{-1/2} \left\| (\Sigma_2 + \lambda_2 I)^{1/2} (\bar{\Sigma}_{2,m} + \lambda_2 I)^{-1} (\Sigma_2 + \lambda_2 I)^{1/2} \right\|_{op} \\
&\times \Big(\left\| (\Sigma_2 + \lambda_2 I)^{-1/2} (\bar{\Sigma}_{2,m} - \Sigma_2)\varphi_{\lambda_2} \right\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \lambda_2^{-1/2} \|\bar{g}_{2,m} - g_2\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \Big),
\end{aligned}$$

The first term is bounded by Lemma (12.45),

$$\left\|(\Sigma_2+\lambda_2 I)^{1/2}(\bar{\Sigma}_{2,m}+\lambda_2 I)^{-1}(\Sigma_2+\lambda_2 I)^{1/2}\right\|_{op} \leq 3,$$

for $m \geq 8\kappa^4 \log(2/\delta) g_{\lambda_2}\lambda_2^{-1}$ with probability at least $1-\delta$ for all $\delta \in (0,1)$. To bound the remaining term, we wish to apply Theorem (12.49) with $\mathcal{H} = \mathcal{H}_{\mathcal{ZA}}$. Consider the measurable map $\xi : \mathcal{W} \times \mathcal{A} \to \mathcal{H}_{\mathcal{ZA}}$ defined by

$$\xi(w,a) := (\Sigma_2+\lambda_2 I)^{-1/2}\langle \varphi_{\lambda_2}, \mu_{Z|W,A}(w,a)\otimes\phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}}\mu_{Z|W,A}(w,a)\otimes\phi_{\mathcal{A}}(a),$$

inducing random variables such that

$$\frac{1}{m}\sum_{i=1}^{m}(\xi(\tilde{w}_i,\tilde{a}_i)-\mathbb{E}[\xi(W,A)]) = (\Sigma_2+\lambda_2 I)^{-1/2}(\bar{\Sigma}_{2,m}-\Sigma_2)\varphi_{\lambda_2}.$$

By Assumption (12.1), Lemma (12.15) and Cauchy-Schwarz inequality,

$$|\langle \varphi_{\lambda_2}, \mu_{Z|W,A}(w,a)\otimes\phi_{\mathcal{A}}(a)\rangle_{\mathcal{H}_{\mathcal{ZA}}}| \leq \kappa^2\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}.$$

We can now bound the q-th moment of ξ, for $q \geq 2$,

$$\begin{aligned}
\mathbb{E}\,\|\xi(W,A)\|^q_{\mathcal{H}_{\mathcal{ZA}}} &\leq \left(\kappa^2\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\right)^q \mathbb{E}\left\|(\Sigma_2+\lambda_2 I)^{-1/2}\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\right\|^q_{\mathcal{H}_{\mathcal{ZA}}} \\
&\leq \left(\kappa^2\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\right)^q \left(\frac{\kappa^2}{\sqrt{\lambda_2}}\right)^{q-2} \mathbb{E}\left\|(\Sigma_2+\lambda_2 I)^{-1/2}\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\right\|^2_{\mathcal{H}_{\mathcal{ZA}}} \\
&= \left(\kappa^2\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\right)^q \left(\frac{\kappa^2}{\sqrt{\lambda_2}}\right)^{q-2} \mathcal{N}(\lambda_2) \\
&\leq \frac{1}{2}q!\left(\frac{\kappa^4}{\sqrt{\lambda_2}}\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\right)^{q-2}\mathcal{N}(\lambda_2)\kappa^4\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{ZA}}},
\end{aligned}$$

where in the equality, we used Proposition (12.16). An application of Bernstein's inequality from Theorem (12.49) with

$$L = \frac{\kappa^4}{\sqrt{\lambda_2}}\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}, \qquad \sigma^2 = \mathcal{N}(\lambda_2)\kappa^4\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{ZA}}},$$

yields the final bound. □

We will now derive a bound for $\|g_2 - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{ZA}}}$.

Lemma 12.18. *With probability at least $1-\delta$ for $\delta \in (0,1)$ the following bound holds:*

$$\|g_2 - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{ZA}}} \leq 2\kappa^2\sqrt{\frac{2\log(2/\delta)}{m(m-1)}}.$$

Proof. Observe that, by Proposition (12.10),

$$\bar{g}_{2,m} = \frac{1}{m(m-1)}\sum_{\substack{i,j\\ i\neq j}}^{m}\mu_{Z|W,A}(\tilde{w}_j,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i),$$

$$\mathbb{E}_W\mathbb{E}_A[\bar{g}_{2,m}] = \mathbb{E}_W\mathbb{E}_A\left[\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\right] = g_2.$$

Let

$$\xi(W,A) := \mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A).$$

Then, note that

$$\|\xi(W,A)\|_{\mathcal{H}_{\mathcal{ZA}}} \leq \kappa^2 \quad \text{(by Assumption (12.1))}.$$

Now, we apply Corollary (12.48) such that with probability at least $1-\delta$,

$$\|g_2 - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq 2\kappa^2\sqrt{\frac{2\log(2/\delta)}{m(m-1)}}.$$

□

Theorem 12.19. *Suppose Assumptions (12.1), (12.6-2.) and (12.9-2) hold. Then, with probability at least $1-2\delta$ for all $\delta \in (0,1/2)$, for $m \geq 8\kappa^4 \log(2/\delta) g_{\lambda_2}\lambda_2^{-1}$,*

$$\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq J_2\left(\frac{\log(2/\delta)}{\sqrt{\lambda_2}}\left(\sqrt{\frac{1}{m}\left(\frac{1}{\lambda_2^{p_2}} + \frac{1}{m\lambda_2}\right)} + \sqrt{\frac{1}{m(m-1)\lambda_2}}\right) + \lambda_2^{\frac{\beta_2-1}{2}}\right),$$

where J_2 is a constant depending on κ, β_2, B_2, D.

Proof. Combining the bounds in Lemma (12.13), Lemma (12.17) and Lemma (12.18) with a union bound, we obtain that with probability at least $1-2\delta$,

$$\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq \mathring{J}\left(\frac{\log(2/\delta)}{\sqrt{\lambda_2}}\left(\sqrt{\frac{1}{m}\left(\mathcal{N}(\lambda_2)\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \frac{\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}}{m\lambda_2}\right)} + \sqrt{\frac{1}{m(m-1)\lambda_2}}\right) + \lambda_2^{\frac{\beta_2-1}{2}}\right),$$

where $\mathring{J}$ is a constant depending on κ, B_2. Under Assumption (12.9-2), using Proposition (12.16), there is a constant $D > 0$ such that $\mathcal{N}(\lambda_2) \leq D\lambda_2^{-p_2}$. Furthermore, under Assumption (12.6-2.),

$$\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = \|\Sigma_2^{\frac{\beta_2-1}{2}}\Sigma_2^{-\frac{\beta_2-1}{2}}\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq \kappa^{\frac{\beta_2-1}{2}}B_2.$$

□

Next, we will derive a bound for $\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|$.

Lemma 12.20. *Under the same assumptions as Theorem (12.11), with probability at least $1-\delta$ for $\delta \in (0,1)$, the following bound holds*

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\| \leq \frac{1}{\lambda_2}\kappa^3 r_1(\delta,n,\beta_1,p_1) + \frac{1}{\lambda_2}\left(\kappa^6 r_1(\delta,n,\beta_1,p_1)^2 + 2B_1\kappa^{6+\frac{\beta_1-1}{2}}r_1(\delta,n,\beta_1,p_1)\right)\|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}.$$

Proof.

$$\begin{aligned}
&\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{g}_{2,m} - \bar{g}_{2,m} + \bar{g}_{2,m}) - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&= \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{g}_{2,m} - \bar{g}_{2,m}) + (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{2,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&= \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{g}_{2,m} - \bar{g}_{2,m}) + (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{2,m} - (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{2,m} + \lambda_2 I)\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&= \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{g}_{2,m} - \bar{g}_{2,m}) + (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{2,m} - (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{2,m}\bar{\varphi}_{\lambda_2,m} + \underbrace{\lambda_2\bar{\varphi}_{\lambda_2,m}}_{\bar{g}_{2,m} - \bar{\Sigma}_{2,m}\bar{\varphi}_{\lambda_2,m}})\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&= \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{g}_{2,m} - \bar{g}_{2,m}) + (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{2,m} - (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{2,m}\\
&- (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{2,m}\bar{\varphi}_{\lambda_2,m} - \bar{\Sigma}_{2,m}\bar{\varphi}_{\lambda_2,m})\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&= \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{g}_{2,m} - \bar{g}_{2,m}) - (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{2,m} - \bar{\Sigma}_{2,m})\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&\leq \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\|_{op}\|\hat{g}_{2,m} - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \|(\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\|_{op}\|\hat{\Sigma}_{2,m} - \bar{\Sigma}_{2,m}\|_{op}\|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\\
&\leq \lambda_2^{-1}\left(\|\hat{g}_{2,m} - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \|\hat{\Sigma}_{2,m} - \bar{\Sigma}_{2,m}\|_{op}\|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\right).
\end{aligned}$$

We have two terms to bound. First, we observe that

$$\begin{aligned}
\|\hat{g}_{2,m} - \bar{g}_{2,m}\|_{\mathcal{H}_{\mathcal{ZA}}} &= \left\| \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ j\neq i}} \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) - \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ j\neq i}} \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{ZA}}} \\
&= \left\| \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ j\neq i}} (\hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{ZA}}} \\
&\leq \frac{\kappa}{m(m-1)} \sum_{\substack{i,j=1 \\ j\neq i}} \left\| \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} \quad \text{(by Assumption (12.1))} \\
&\leq \kappa^3 r_1(\delta, n, \beta_1, p_1) \quad \text{(with probability } 1-\delta \text{, Corollary (12.12)).}
\end{aligned}$$

For the second component, we note that for $i = 1, \ldots, m$,

$$\begin{aligned}
\xi_i &:= (\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \otimes (\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) - (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \otimes (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \\
&= ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \otimes ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \\
&+ ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \otimes (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \\
&+ (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)) \otimes ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i)),
\end{aligned}$$

and therefore, under Assumption (12.1), by the triangular inequality and by Corollary (12.12),

$$\begin{aligned}
\|\xi_i\|_{op} &\leq \kappa^2 \left\| \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|^2_{\mathcal{H}_{\mathcal{Z}}} + 2\kappa^2 \left\| \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} \left\| \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} \\
&\leq \kappa^6 r_1(\delta, n, \beta_1, p_1)^2 + 2\kappa^4 \left\| \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} r_1(\delta, n, \beta_1, p_1).
\end{aligned}$$

with probability at least $1-\delta$. We also note that under Assumptions (5.1), (12.1) and (12.6-1), for all $(w, a) \in \mathcal{W} \times \mathcal{A}$,

$$\begin{aligned}
\|\mu_{Z|W,A}(w, a)\|_{\mathcal{H}_{\mathcal{Z}}} &= \|C_{Z|W,A}(\phi_{\mathcal{W}}(w) \otimes \phi_{\mathcal{A}}(a))\|_{\mathcal{H}_{\mathcal{Z}}} \\
&\leq \|C_{Z|W,A}\|_{S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})} \|\phi_{\mathcal{W}}(w) \otimes \phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{WA}}} \\
&\leq \kappa^2 \|C_{Z|W,A}\|_{S_2(\mathcal{H}_{\mathcal{WA}}, \mathcal{H}_{\mathcal{Z}})} \\
&\leq B_1 \kappa^{2 + \frac{\beta_1 - 1}{2}}.
\end{aligned}$$

Finally, we obtain that with probability at least $1-\delta$,

$$\begin{aligned}
\|\hat{\Sigma}_{2,m} - \bar{\Sigma}_{2,m}\|_{op} &= \left\| \frac{1}{m} \sum_{i=1}^{m} \xi_i \right\|_{op} \leq \frac{1}{m} \sum_{i=1}^{m} \|\xi_i\|_{op} \\
&\leq \kappa^6 r_1(\delta, n, \beta_1, p_1)^2 + 2B_1 \kappa^{6 + \frac{\beta_1 - 1}{2}} r_1(\delta, n, \beta_1, p_1). \qquad (31)
\end{aligned}$$

□

The following Theorem provides convergence rates in RKHS norm for the estimation of the bridge solution with minimum RKHS norm.

Theorem 12.21. *Suppose Assumptions (5.1-1 & 2), (12.1), (12.6-1 & 2) and (12.9-1 & 2) hold and set* $\lambda_1 = \Theta\left(n^{-\frac{1}{\beta_1 + p_1}}\right)$ *and* $n = m^{\iota \frac{\beta_1 + p_1}{\beta_1 - 1}}$ *where* $\iota > 0$. *Then,*

i. If $\iota \leq \frac{\beta_2 + 1}{\beta_2 + p_2}$ *then* $\|\hat{\varphi}_{\lambda_2, m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} = O_p\left(m^{-\frac{\iota}{2} \frac{\beta_2 - 1}{\beta_2 + 1}}\right)$ *with* $\lambda_2 = \Theta\left(m^{-\frac{\iota}{\beta_2 + 1}}\right)$;

ii. If $\iota \geq \frac{\beta_2 + 1}{\beta_2 + p_2}$ *then* $\|\hat{\varphi}_{\lambda_2, m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} = O_p\left(m^{-\frac{1}{2} \frac{\beta_2 - 1}{\beta_2 + p_2}}\right)$ *with* $\lambda_2 = \Theta\left(m^{-\frac{1}{\beta_2 + p_2}}\right)$.

Proof. Let us abbreviate $r_1(n) = r_1(\delta, n, \beta_1, p_1)$. From Lemma (12.20), we obtain with high probability that

$$\begin{aligned}\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} &\le \|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \frac{\bar{J}}{\lambda_2} r_1(n)\left(1 + (r_1(n)+1)\|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\right) \\ &\le \|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \frac{\bar{J}}{\lambda_2} r_1(n) + \frac{\bar{J}}{\lambda_2}(1 + r_1(n)) r_1(n)(\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}),\end{aligned}$$

where $\bar{J}$ is a constant depending on κ, B_1, β_1. Furthermore, from Theorem (12.19),

$$\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = O_p\Bigg(r_2(m)\Bigg) \quad \text{with} \quad r_2(m) = \frac{1}{\sqrt{\lambda_2}}\left(\sqrt{\frac{1}{m}\left(\frac{1}{\lambda_2^{p_2}} + \frac{1}{m\lambda_2}\right)} + \sqrt{\frac{1}{m(m-1)}}\right) + \lambda_2^{\frac{\beta_2-1}{2}}.$$

as long as $m \ge 8\kappa^4 \log(2/\delta) g_{\lambda_2}\lambda_2$. Note that the term $[m(m-1)]^{-1/2}$ is of faster order and can be removed. Putting it together, we obtain,

$$\begin{aligned}\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} &= O_p\left(\sqrt{\frac{1}{m\lambda_2^{1+p_2}}\left(1 + \frac{1}{m\lambda_2^{1-p_2}}\right)} + \lambda_2^{\frac{\beta_2-1}{2}} + \frac{r_1(n)}{\lambda_2} + \frac{r_1(n)(1 + r_1(n))(\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\| + \|\bar{\varphi}_0\|)}{\lambda_2}\right) \\ &= O\left(\sqrt{\frac{1}{m\lambda_2^{1+p_2}}\left(1 + \frac{1}{m\lambda_2^{1-p_2}}\right)} + \lambda_2^{\frac{\beta_2-1}{2}} + \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}}\right),\end{aligned}$$

where in the second equation we removed the last term that is of faster order with respect to $r_1(n)\lambda_2^{-1}$ and we plugged the expression for $r_1(n)$ given in Theorem (12.11). We are now ready to prove cases i. and ii. For each choice of λ_2 we are required to check the condition $g_{\lambda_2}\lambda_2^{-1}m^{-1} = O(1)$ where $g_{\lambda_2} = \log\left(2e\mathcal{N}(\lambda_2)\frac{\|\Sigma_2\|_{op}+\lambda_2}{\|\Sigma_2\|_{op}}\right)$. Let us fix a lower bound $0 < c \le 1$ with $c \le \|\Sigma_2\|_{op}$. Since $\lambda_2 \to 0$ we choose $m_0 \ge 1$ such that $\lambda_2 \le c \le \min\{1, \|\Sigma_2\|_{op}\}$ for all $m \ge m_0$. We get for $m \ge m_0$, by Proposition (12.16),

$$\begin{aligned}\frac{g_{\lambda_2}}{m\lambda_2} &= \frac{1}{m\lambda_2} \cdot \log\left(2e\mathcal{N}(\lambda_2)\frac{\|\Sigma_2\|_{op} + \lambda_2}{\|\Sigma_2\|_{op}}\right) \\ &\le \frac{1}{m\lambda_2} \cdot \log\left(4De\lambda_2^{-p_2}\right) \\ &= \frac{\log(4De)}{m\lambda_2} + \frac{p_2 \log \lambda_2^{-1}}{m\lambda_2}.\end{aligned}$$

Therefore, to check $g_{\lambda_2}\lambda_2^{-1}m^{-1} = O(1)$, it is sufficient to check $\lambda_2^{-1}m^{-1}\log\lambda_2^{-1} = O(1)$.

Case i. Let $n = m^{\iota\frac{\beta_1+p_1}{\beta_1-1}}$ with $\iota \le \frac{\beta_2+1}{\beta_2+p_2}$ and $\lambda_2 = m^{-\frac{\iota}{\beta_2+1}}$. We first check $\lambda_2^{-1}m^{-1}\log\lambda_2^{-1} = O(1)$, with this choice of λ_2, we have,

$$\frac{\log\lambda_2^{-1}}{m\lambda_2} = \frac{\iota}{\beta_2+1}\frac{\log(m)}{m}m^{\frac{\iota}{\beta_2+1}} = \frac{\iota}{\beta_2+1}\frac{\log(m)}{m^{1-\frac{\iota}{\beta_2+1}}}.$$

As $\iota \le (\beta_2+1)(\beta_2+p_2)^{-1}$ implies $1 - \iota(\beta_2+1)^{-1} > 0$, we have $\log(\lambda_2^{-1})/(m\lambda_2) \to 0$, as $m \to \infty$. Next note that with this choice of λ_2, we have,

$$\lambda_2^{\frac{\beta_2-1}{2}} = m^{-\frac{\iota}{2}\frac{\beta_2-1}{\beta_2+1}} = \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}}.$$

Furthermore,

$$\frac{1}{m\lambda_2^{1+p_2}} \le \lambda_2^{\beta_2-1} \iff \iota \le \frac{\beta_2+1}{\beta_2+p_2},$$

and

$$\frac{1}{m\lambda_2^{1-p_2}} \le \frac{1}{m\lambda_2^{1+p_2}} \le 1.$$

Therefore,

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = O\left(\lambda_2^{\frac{\beta_2-1}{2}}\right) = O\left(m^{-\frac{\iota}{2}\frac{\beta_2-1}{\beta_2+1}}\right).$$

Case ii. Let $n = m^{\iota \frac{\beta_1+p_1}{\beta_1-1}}$ with $\iota \geq \frac{\beta_2+1}{\beta_2+p_2}$ and $\lambda_2 = m^{-\frac{1}{\beta_2+p_2}}$. We first check $\lambda_2^{-1} m^{-1} \log \lambda_2^{-1} = O(1)$, with this choice of λ_2, we have,

$$\frac{\log \lambda_2^{-1}}{m\lambda_2} = \frac{1}{\beta_2+p_2} \frac{\log(m)}{m} m^{\frac{1}{\beta_2+p_2}} = \frac{1}{\beta_2+p_2} \frac{\log(m)}{m^{1-\frac{1}{\beta_2+p_2}}}.$$

As $\beta_2 + p_2 > 1$, we have $\log(\lambda_2^{-1})/(m\lambda_2) \to 0$, as $m \to \infty$. Next note that with this choice of λ_2, we have,

$$\frac{1}{\lambda_2}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}} = \lambda_2^{\frac{\beta_2-1}{2}} \left(\lambda_2^{-\frac{\beta_2+1}{2}} \left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}}\right),$$

and

$$\lambda_2^{-\frac{\beta_2+1}{2}} \left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}} = \sqrt{m}^{\frac{\beta_2+1}{\beta_2+p_2}} \left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}} \leq 1 \iff n \geq m^{\frac{\beta_1+p_1}{\beta_1-1}\frac{\beta_2+1}{\beta_2+p_2}}.$$

Therefore, $\lambda_2^{-1}\sqrt{n}^{-\frac{\beta_1-1}{\beta_1+p_1}} \leq \lambda_2^{\frac{\beta_2-1}{2}}$ since we have assumed $n = m^{\iota\frac{\beta_1+p_1}{\beta_1-1}} \geq m^{\frac{\beta_1+p_1}{\beta_1-1}\frac{\beta_2+1}{\beta_2+p_2}}$. Furthermore,

$$\frac{1}{m\lambda_2^{1+p_2}} = \left(\frac{1}{m}\right)^{\frac{\beta_2-1}{\beta_2+p_2}} = \lambda_2^{\beta_2-1},$$

and

$$\frac{1}{m\lambda_2^{1-p_2}} = \left(\frac{1}{m}\right)^{\frac{\beta_2-1+2p_2}{\beta_2+p_2}} \leq \frac{1}{m\lambda_2^{1+p_2}}.$$

Therefore,

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = O\left(\lambda_2^{\frac{\beta_2-1}{2}}\right) = O\left(\sqrt{m}^{-\frac{\beta_2-1}{\beta_2+p_2}}\right).$$

□

12.1.5 Third-Stage Regression Consistency Results

The following theorem is obtained from Li et al. (2024) that provides convergence guarantees for vector-valued regression with Tikhonov regularization.

Theorem 12.22 (Theorem 3 Li et al. (2024))**.** *Suppose Assumptions (5.1-3), (12.1), (12.6-3) and (12.9-3) hold and take $\lambda_3 = \Theta\left(t^{-\frac{1}{\beta_3+p_3}}\right)$. There is a constant $J_3 > 0$ independent of $t \geq 1$ and $\delta \in (0,1)$ such that*

$$\left\|\hat{C}_{YZ|A} - C_{YZ|A}\right\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})} \leq J_3 \log(5/\delta)\left(\frac{1}{\sqrt{t}}\right)^{\frac{\beta_3-1}{\beta_3+p_3}} =: r_3(\delta, t, \beta_3, p_3),$$

is satisfied for sufficiently large $t \geq 1$ with probability at least $1-\delta$.

Proof. To apply Theorem 3 Li et al. (2024), we prove that the following noise condition is satisfied, for $q \geq 2$, P_A−almost surely,

$$\mathbb{E}[\|Y\phi_{\mathcal{Z}}(Z) - \Psi(A)\|^q_{\mathcal{H}_{\mathcal{Z}}} \mid A] \leq \frac{1}{2} q! \tilde{\sigma}^2 \tilde{R}^{q-2}, \tag{32}$$

for some $\tilde{\sigma} > 0$ and $\tilde{R} > 0$. We first notice that the Bernstein condition given by Assumption 12.1 Eq. (29) implies that conditionally on A, $|Y - \mathbb{E}[Y \mid A]|$ is sub-exponential (Section 1.4 Buldygin and Kozachenko, 2000), which then implies sub-exponentiality of $|Y|$ given A. As, $\|\phi_{\mathcal{Z}}(Z)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \kappa$ is almost surely bounded, we obtain that $\|Y\phi_{\mathcal{Z}}(Z)\|_{\mathcal{H}_{\mathcal{Z}}}$ given A is sub-exponential. Finally, Mollenhauer et al. (2022, Remark 4.9 and Appendix A.2) shows that $\|Y\phi_{\mathcal{Z}}(Z)\|_{\mathcal{H}_{\mathcal{Z}}}$ given A being sub-exponential implies Eq. (32). We can then apply Theorem 3 Li et al. (2024) with $\gamma = 1$ which corresponds to the Hilbert-Schmidt norm of $S_2(\mathcal{H}_{\mathcal{A}}, \mathcal{H}_{\mathcal{Z}})$. Similarly to the proof of Theorem 12.11, we can take the smoothness up to $\beta_3 = 3$ as we work with the RKHS norm. □

Corollary 12.23. *Under the same assumptions as Theorem (12.22), with $\lambda_3 = \Theta\left(n^{-\frac{1}{\beta_3+p_3}}\right)$, for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\sup_{a\in\mathcal{A}} \|\hat{\mu}_{YZ|A}(a) - \mu_{YZ|A}(a)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \kappa r_3(\delta, t, \beta_3, p_3).$$

12.1.6 Consistency for Dose-Response Curve

Theorem 12.24. *Suppose Assumptions (5.1), (12.1), (12.6) and (12.9) hold and set* $\lambda_1 = \Theta\left(n^{-\frac{1}{\beta_1+p_1}}\right)$, $\lambda_3 = \Theta\left(t^{-\frac{1}{\beta_3+p_3}}\right)$ *and* $n = m^{\iota\frac{\beta_1+p_1}{\beta_1-1}}$ *where* $\iota > 0$*. Then,*

i. If $\iota \leq \frac{\beta_2+1}{\beta_2+p_2}$ *then* $\|\hat{f}_{ATE} - f_{ATE}\|_\infty = O_p\left(\sqrt{t}^{-\frac{\beta_3-1}{\beta_3+p_3}} + m^{-\frac{\iota}{2}\frac{\beta_2-1}{\beta_2+1}}\right)$ *with* $\lambda_2 = \Theta\left(m^{-\frac{\iota}{\beta_2+1}}\right)$*;*

ii. If $\iota \geq \frac{\beta_2+1}{\beta_2+p_2}$ *then* $\|\hat{f}_{ATE} - f_{ATE}\|_\infty = O_p\left(\sqrt{t}^{-\frac{\beta_3-1}{\beta_3+p_3}} + m^{-\frac{1}{2}\frac{\beta_2-1}{\beta_2+p_2}}\right)$ *with* $\lambda_2 = \Theta\left(m^{-\frac{1}{\beta_2+p_2}}\right)$*.*

Proof. We recall that we showed in Equation (30) that for any $a \in \mathcal{A}$,

$$\begin{aligned}&|\hat{f}_{ATE}(a) - f_{ATE}(a)| \\ &\leq \kappa\left(\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} + \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \|\hat{\varphi}_{\lambda_2,m} - \varphi_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\|\Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}}\right).\end{aligned}$$

Note that, under Assumption (5.1-3), Assumption (12.1) and Assumption (12.6-3),

$$\|\Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} = \|C_{YZ|A}\phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \kappa\|C_{YZ|A}\Sigma_3^{-\frac{\beta_3-1}{2}}\Sigma_3^{\frac{\beta_3-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})} \leq B_3\kappa^{1+\frac{\beta_3-1}{2}} =: \alpha_1.$$

Furthermore, under Assumption (12.6-2), by Proposition 12.7,

$$\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = \|\Sigma_2^{\frac{\beta_2-1}{2}}\Sigma_2^{-\frac{\beta_2-1}{2}}\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq B_2\kappa^{\frac{\beta_2-1}{2}} =: \alpha_2.$$

Therefore,

$$|\hat{f}_{ATE}(a) - f_{ATE}(a)| \leq \kappa\left(\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} + \alpha_2\|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \alpha_1\|\hat{\varphi}_{\lambda_2,m} - \varphi_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}\right).$$

As the first term is faster than the two last terms, plugging the results of Theorem 12.21 and Corollary 12.23, we obtain the final bound. □

12.2 Consistency Results for Conditional Dose-Response

In this section, we present the consistency result for the estimation of f_{ATT}. We note that the first-stage and third-stage regressions are identical to the ones for f_{ATE}. Hence, we first derive the consistency of the second-stage regression of conditional dose-response curve. Then, by incorporating the first and third-stage consistency results for dose-response curve from previous sections, we prove the non-asymptotic uniform consistency guarantees for conditional dose-response curve estimation.

Throughout the whole section $a' \in \mathcal{A}$ is fixed.

12.2.1 Second-Stage Regression Consistency Results

First, we will assume that the problem is well-defined.

Assumption 12.25 (RKHS bridge for ATT)**.** *There exists* $\varphi_0 \in \mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}$ *that is a solution of Equation (2).*

Similarly to ATE we define the minimum norm bridge solution for ATT.

Definition 12.26 (Bridge solution with minimum RKHS norm)**.** *We define*

$$\bar{\varphi}_0 = \underset{\varphi\in\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}{\arg\min}\,\|\varphi\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} \quad s.t. \quad \mathbb{E}[\varphi(Z, A, a')|W, A] = r(W, A, a'),$$

where $r(W, A, a') = \frac{p(W,a')p(a)}{p(W,a)p(a')}$*.*

Definition 12.27. *We define the second stage covariance operator for the conditional dose-response:*

$$\Sigma_4 := \mathbb{E}_{W,A}\left[\left(\left(\mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')\right) \otimes \left(\mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')\right)\right)\right].$$

Proposition 12.28. *Under Assumption (12.25), $\bar{\varphi}_0$ is well-defined and is the unique element of $\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}$ satisfying Equation (2) and such that $\bar{\varphi}_0 \in \operatorname{null}(\Sigma_4)^{\perp}$.*

Proof. Note that the bridge equation, for an element $\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}$, can be written as

$$r(W, A, a') = E[\varphi(Z, A, a')|W, A] = \langle \varphi, \mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a') \rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} = A\varphi,$$

by using the reproducing property and introducing the operator $A : \mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}} \to \mathcal{L}_2(\mathcal{W} \times \mathcal{A}, p_{W,A}), \varphi \mapsto \langle \varphi, \mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a') \rangle_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}$. The rest of the proof follows from the steps of the proof of Proposition (12.5). □

Recall that we have the following loss at the population level:

$$\mathcal{L}^{2SR}(\varphi) = \mathbb{E}[(r(W, A, a') - \mathbb{E}[\varphi(Z, A, a')|W, A])^2].$$

This loss can be equivalently written as

$$\mathcal{L}^{2SR}(\varphi) = \mathbb{E}[\mathbb{E}[\varphi(Z, A, a')|W, A]^2] - 2\mathbb{E}_{W|A'=a'}\mathbb{E}_A[\mathbb{E}[\varphi(Z, A, a')|W, A]] + \text{const}.$$

We introduce the regularized version of the population loss, for $\lambda_2 > 0$,

$$\mathcal{L}^{2SR}_{\lambda_2}(\varphi) = \mathcal{L}^{2SR}(\varphi) + \lambda_2 \|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}.$$

We modify Assumptions (12.6) and (12.9) for ATT as follows.

Assumption 12.29. *We assume that the following condition holds: there exists a constant $B_4 < \infty$ such that for a given $\beta_4 \in (1, 3]$,*

$$\|\Sigma_4^{-\frac{\beta_4 - 1}{2}} \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} \leq B_4.$$

Assumption 12.30. *We assume that the following condition holds: let $(\lambda_{4,i})_{i \geq 1}$ be the eigenvalues of Σ_4, for some constant $c_4 > 0$ and parameter $p_4 \in (0, 1]$ and for all $i \geq 1$,*

$$\lambda_{4,i} \leq c_4 i^{-1/p_4}.$$

Let us also introduce g_4 defined as

$$g_4 = \mathbb{E}_{W|A'=a'}\mathbb{E}_A[\mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')].$$

Proposition 12.31. *g_4 can be equivalently written as*

$$g_4 = \mathbb{E}[r(W, A, a')\mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')].$$

Furthermore, for any element $\varphi_0 \in \mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}$ solution to the bridge function equation, we have $g_4 = \Sigma_4 \varphi_0$. Finally,

$$\varphi_{\lambda_2} := \underset{\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}{\arg\min} \mathcal{L}^{2SR}_{\lambda_2}(\varphi) = (\Sigma_4 + \lambda_2 Id_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}})^{-1} g_4.$$

Proof. The proof follows the same steps of the proof of Proposition (12.10) □

Combined with Proposition (12.28), we obtain the following result.

Proposition 12.32. *Under Assumption (12.25), $\bar{\varphi}_0 = \Sigma_4^{\frac{\beta_4 - 1}{2}} \Sigma_4^{-\frac{\beta_4 - 1}{2}} \bar{\varphi}_0$.*

Now, we introduce the empirical version of the loss function. Define

$$\bar{\mathcal{L}}^{2SR}_m(\varphi) = \frac{1}{m} \sum_{i=1}^{m} \left\langle \varphi, \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \right\rangle^2$$

$$- \frac{2}{m} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \left\langle \varphi, \theta_i \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a') \right\rangle + \lambda_2 \|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}},$$

$$\begin{aligned} \hat{\mathcal{L}}^{2SR}_m(\varphi) &= \frac{1}{m} \sum_{i=1}^{m} \left\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \right\rangle^2 \\ &\quad - \frac{2}{m} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \left\langle \varphi, \theta_i \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a') \right\rangle + \lambda_2 \|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}. \end{aligned}$$

where $\theta_i = [(\boldsymbol{K}_{\tilde{A}\tilde{A}} + m\zeta \boldsymbol{I})^{-1} \boldsymbol{K}_{\tilde{A}a'}]_i$. To write down the minimizers of these loss functions, we introduce the following sample based operators

$$\begin{aligned} \bar{\Sigma}_{4,m} &= \frac{1}{m} \sum_{i=1}^{m} \Big(\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big) \otimes \Big(\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big), \\ \bar{g}_{4,m} &= \frac{1}{m} \sum_{\substack{i,j \\ i \neq j}}^{m} \theta_i \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \\ \hat{\Sigma}_{4,m} &= \frac{1}{m} \sum_{i=1}^{m} \Big(\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big) \otimes \Big(\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') \Big), \\ \hat{g}_{4,m} &= \frac{1}{m} \sum_{\substack{i,j \\ i \neq j}}^{m} \theta_i \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'). \end{aligned}$$

We can observe that the minimizers of the objective functions are given by

$$\begin{aligned} \varphi_{\lambda_2} &= (\Sigma_4 + \lambda_2 I)^{-1} g_4 = \underset{\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}}}{\arg\min}\, \mathcal{L}^{2SR}_{\lambda_2}(\varphi), \\ \bar{\varphi}_{\lambda_2,m} &= (\bar{\Sigma}_{4,m} + \lambda_2 I)^{-1} \bar{g}_{4,m} = \underset{\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}}}{\arg\min}\, \bar{\mathcal{L}}^{2SR}_m(\varphi), \\ \hat{\varphi}_{\lambda_2,m} &= (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1} \hat{g}_{4,m} = \underset{\varphi \in \mathcal{H}_{\mathcal{Z}\mathcal{A}}}{\arg\min}\, \hat{\mathcal{L}}^{2SR}_m(\varphi). \end{aligned}$$

$\hat{\varphi}_{\lambda_2,m}$ is the final estimator presented in Section (4.2). To show the convergence of $\hat{\varphi}_{\lambda_2,m}$ to the minimum norm bridge $\bar{\varphi}_0$ introduced in Definition (12.26), we will consider the following decomposition,

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} \leq \|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} + \|\bar{\varphi}_{\lambda_2,m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} + \|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}.$$

First, consider an upper bound for $\|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}}$

Lemma 12.33. *Suppose that Assumption (12.29) holds with parameter $\beta_4 \in (1, 3]$. Then, for any $\lambda_2 > 0$,*

$$\|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} \leq B_4 \lambda_2^{\frac{\beta_4 - 1}{2}}.$$

Proof. We saw in Proposition (12.31) that

$$\varphi_{\lambda_2} = (\Sigma_4 + \lambda_2 \operatorname{Id})^{-1} g_4 = (\Sigma_4 + \lambda_2 \operatorname{Id})^{-1} \Sigma_4 \bar{\varphi}_0 = \bar{\varphi}_0 - \lambda_2 (\Sigma_4 + \lambda_2 \operatorname{Id})^{-1} \bar{\varphi}_0.$$

Therefore, under Assumption (12.29),

$$\|\varphi_{\lambda_2} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} = \lambda_2 \left\| (\Sigma_4 + \lambda_2 \operatorname{Id})^{-1} \bar{\varphi}_0 \right\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} \leq B_4 \lambda_2 \left\| (\Sigma_4 + \lambda_2 \operatorname{Id})^{-1} \Sigma_4^{\frac{\beta_4 - 1}{2}} \right\|_{op},$$

where we used Proposition (12.32). Note that, by Lemma (12.46),

$$\left\|\left(\Sigma_4 + \lambda_2 \operatorname{Id}\right)^{-1} \Sigma_4^{\frac{\beta_4 - 1}{2}}\right\|_{op} = \sup_{i \geq 1} \frac{\lambda_{4,i}^{\frac{\beta_4-1}{2}}}{\lambda_{4,i} + \lambda_2} \leq \lambda_2^{\frac{\beta_4 - 1}{2} - 1},$$

as long as $\frac{\beta_4 - 1}{2} \in (0, 1]$, i.e. $\beta_4 \in (1, 3]$. By merging the bounds, we obtain the final result. □

Lemma 12.34. *For any $\lambda_2 > 0$, $\|\varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}$.*

Proof. We saw in Proposition (12.31) that

$$\|\varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZAA}}} = \left\|\left(\Sigma_4 + \lambda_2 \operatorname{Id}\right)^{-1} \Sigma_4 \bar{\varphi}_0\right\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq \left\|\left(\Sigma_4 + \lambda_2 \operatorname{Id}\right)^{-1} \Sigma_4\right\|_{op} \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}.$$

□

Similar to the consistency result for the dose-response curve, our convergence proof uses both the Hoeffding concentration inequality (Corollary 12.48) and the Bernstein concentration inequality (Theorem 12.49) for Hilbert space-valued random variables. We also define the effective dimension for the stage 2 error: for $\lambda_2 > 0$, $\mathcal{N}(\lambda_2) := \operatorname{Tr}((\Sigma_4 + \lambda_2 \operatorname{Id})^{-1} \Sigma_4)$ Caponnetto and De Vito (2007).

Proposition 12.35 (Lemma 11 & Lemma 13 Fischer and Steinwart (2020))**.** *Under Assumption (12.30), there is a constant $D > 0$ such that the following inequality is satisfied, for $\lambda_2 > 0$, $\mathcal{N}(\lambda_2) \leq D \lambda_2^{-p_4}$. Furthermore, we have the equality,*

$$\mathbb{E}\left[\left\|(\Sigma_4 + \lambda_2 \operatorname{Id})^{-1/2} \mu_{Z|W,A}(W, A) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')\right\|_{\mathcal{H}_{\mathcal{ZAA}}}^2\right] = \mathcal{N}(\lambda_2).$$

Lemma 12.36. *Let us introduce $g_{\lambda_2} = \log\left(2e\mathcal{N}(\lambda_2)\frac{\|\Sigma_4\|_{op} + \lambda_2}{\|\Sigma_4\|_{op}}\right)$. Suppose Assumption (12.1) holds. Then, with probability at least $1 - \delta$ for all $\delta \in (0, 1)$, for $m \geq 8\kappa^6 \log(2/\delta) g_{\lambda_2} \lambda_2^{-1}$,*

$$\|\bar{\varphi}_{\lambda_2, m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq \frac{3}{\sqrt{\lambda_2}}\left(\log(2/\delta)\sqrt{\frac{32}{m}\left(\mathcal{N}(\lambda_2)\kappa^6 \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}^2 + \frac{\kappa^{12}\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}^2}{m\lambda_2}\right)} + \frac{\|g_4 - \bar{g}_{4,m}\|_{\mathcal{H}_{\mathcal{ZAA}}}}{\sqrt{\lambda_2}}\right).$$

Proof. We decompose the error as

$$\begin{aligned}
\|\bar{\varphi}_{\lambda_2, m} - \varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZAA}}} &= \left\|(\bar{\Sigma}_{4,m} + \lambda_2 I)^{-1}\big(\bar{g}_{4,m} - (\bar{\Sigma}_{4,m} + \lambda_2 I)\varphi_{\lambda_2}\big)\right\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&\leq \left\|(\Sigma_4 + \lambda_2 I)^{-1/2}\right\|_{op} \left\|(\Sigma_4 + \lambda_2 I)^{1/2}(\bar{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\Sigma_4 + \lambda_2 I)^{1/2}\right\|_{op} \\
&\quad \times \Big\|(\Sigma_4 + \lambda_2 I)^{-1/2}\big(\bar{g}_{4,m} - \bar{\Sigma}_{4,m}\varphi_{\lambda_2} - \underbrace{\lambda_2 \varphi_{\lambda_2}}_{= g_4 - \Sigma_4 \varphi_{\lambda_2}}\big)\Big\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&\leq \lambda_2^{-1/2} \left\|(\Sigma_4 + \lambda_2 I)^{1/2}(\bar{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\Sigma_4 + \lambda_2 I)^{1/2}\right\|_{op} \\
&\quad \times \left(\|(\Sigma_4 + \lambda_2 I)^{-1/2}(\bar{\Sigma}_{4,m} - \Sigma_4)\varphi_{\lambda_2}\|_{\mathcal{H}_{\mathcal{ZAA}}} + \lambda_2^{-1/2}\|\bar{g}_{4,m} - g_4\|_{\mathcal{H}_{\mathcal{ZAA}}}\right),
\end{aligned}$$

where $\|.\|_{op}$ denotes the *operator norm.* In the same fashion as in Lemma (12.45), we obtain that

$$\left\|(\Sigma_4 + \lambda_2 I)^{1/2}(\bar{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\Sigma_4 + \lambda_2 I)^{1/2}\right\|_{op} \leq 3,$$

for $m \geq 8\kappa^6 \log(2/\delta) g_{\lambda_2} \lambda_2^{-1}$ with probability at least $1 - \delta$ for all $\delta \in (0, 1)$. To bound the remaining term, we wish to apply Theorem (12.49) with $\mathcal{H} = \mathcal{H}_{\mathcal{ZAA}}$. Consider the measurable map $\xi : \mathcal{W} \times \mathcal{A} \rightarrow \mathcal{H}_{\mathcal{ZAA}}$ defined by

$$\xi(w, a) := (\Sigma_4 + \lambda_2 I)^{-1/2}\langle \varphi_{\lambda_2}, \mu_{Z|W,A}(w, a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZA}}} \mu_{Z|W,A}(w, a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a'),$$

inducing random variables such that

$$\frac{1}{m}\sum_{i=1}^{m}\left(\xi(\tilde{w}_i,\tilde{a}_i)-\mathbb{E}[\xi(W,A)]\right)=(\Sigma_4+\lambda_2 I)^{-1/2}(\bar{\Sigma}_{4,m}-\Sigma_4)\varphi_{\lambda_2}.$$

By Assumption (12.1), Lemma (12.34) and Cauchy-Schwarz inequality,

$$|\langle\varphi_{\lambda_2},\mu_{Z|W,A}(w,a)\otimes\phi_{\mathcal{A}}(a)\otimes\phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}}|\leq\kappa^3\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}.$$

We can now bound the q-th moment of ξ, for $q\geq 2$,

$$\begin{aligned}\mathbb{E}\,\|\xi(W,A)\|^q_{\mathcal{H}_{\mathcal{ZA}}}&\leq\left(\kappa^3\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}\right)^q\mathbb{E}\left\|(\Sigma_4+\lambda_2 I)^{-1/2}\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\otimes\phi_{\mathcal{A}}(a')\right\|^q_{\mathcal{H}_{\mathcal{ZAA}}}\\&\leq\left(\kappa^3\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}\right)^q\left(\frac{\kappa^3}{\sqrt{\lambda_2}}\right)^{q-2}\mathbb{E}\left\|(\Sigma_4+\lambda_2 I)^{-1/2}\mu_{Z|W,A}(W,A)\otimes\phi_{\mathcal{A}}(A)\otimes\phi_{\mathcal{A}}(a')\right\|^2_{\mathcal{H}_{\mathcal{ZAA}}}\\&=\left(\kappa^3\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}}\right)^q\left(\frac{\kappa^3}{\sqrt{\lambda_2}}\right)^{q-2}\mathcal{N}(\lambda_2)\\&\leq\frac{1}{2}q!\left(\frac{\kappa^6}{\sqrt{\lambda_2}}\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}}\right)^{q-2}\mathcal{N}(\lambda_2)\kappa^6\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{ZA}}},\end{aligned}$$

where in the equality, we used Proposition (12.35). An application of Bernstein's inequality from Theorem (12.49) with

$$L=\frac{\kappa^6}{\sqrt{\lambda_2}}\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}},\qquad\sigma^2=\mathcal{N}(\lambda_2)\kappa^6\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{ZA}}},$$

yields the final bound. □

We will now derive a bound for $\|g_4-\bar{g}_{4,m}\|_{\mathcal{H}_{\mathcal{ZA}}}$. We need the following assumption.

Assumption 12.37. *Suppose that there exists $C_{W|A}\in S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{W}})$ such that $\mathbb{E}[\phi_{\mathcal{W}}(W)|A]=C_{W|A}\phi_{\mathcal{A}}(A)$.*

Since we also need the consider the variables θ_i in deriving this bound, we will introduce an intermediate operator as follows:

$$\begin{aligned}\tilde{g}_4&=\frac{1}{m}\sum_{j=1}^{m}C_{Z|W,A}\Big(\mathbb{E}_{W|A=a'}[\phi_{\mathcal{W}}(W)]\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\Big)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\\&=\frac{1}{m}\sum_{j=1}^{m}C_{Z|W,A}\Big(C_{W|A}\phi_{\mathcal{A}}(a')\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\Big)\otimes\phi_{\mathcal{A}}(\tilde{a}_j)\otimes\phi_{\mathcal{A}}(a')\end{aligned}$$

where $C_{W|A}\in S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{W}})$ is the conditional mean operator, i.e., $\mathbb{E}[\phi_{\mathcal{W}}(W)|A=a]=C_{W|A}\phi_{\mathcal{A}}(a)$. Recall that, in our conditional dose-response algorithm, we estimate this operator using second-stage data. Therefore, we need the following bound for the estimation error of $C_{W|A}$.

Assumption 12.38. *We assume that the following condition holds: there exist a constant $B_5<\infty$ such that for a given $\beta_5\in(1,3]$,*

$$\|C_{W|A}\Sigma_3^{-\frac{\beta_5-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{W}})}\leq B_5.$$

Theorem 12.39 (Theorem 2 Li et al. (2022))**.** *Suppose Assumptions (12.1), (12.9-3), (12.37), (12.38), hold, and take $\zeta=\Theta\left(m^{-\frac{1}{\beta_5+p_3}}\right)$. Then, there is a constant $J_5>0$ independent of $m\geq 1$ and $\delta\in(0,1)$ such that*

$$\|\hat{C}_{W|A}-C_{W|A}\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{W}})}\leq J_5\log(4/\delta)\left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5-1}{\beta_5+p_3}}$$

is satisfied for sufficiently large $m\geq 1$ with probability at least $1-\delta$.

Lemma 12.40. *With probability at least $1-\delta$ for $\delta \in (0,1)$ the following bound holds:*

$$\|\bar{g}_{4,m} - g_4\| \leq \kappa^4 J_5 \log(4/\delta) \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} \left\|C_{Z|W,A}\right\| + 2\kappa^3 \sqrt{\frac{2\log(2/\delta)}{m}}$$

Proof. Let

$$\begin{aligned} \xi(A) &= C_{Z|W,A}\left(\mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(A)\right) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a') \\ \xi_i &= C_{Z|W,A}\left(\mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(a_i)\right) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a') \end{aligned}$$

Note that $\mathbb{E}[\xi_i] = g_4$. Furthermore, we observe that

$$\begin{aligned} \|\xi(A)\| &= \|C_{Z|W,A}\left(\mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(A)\right) \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')\| \\ &= \|\mathbb{E}_{W|A=a'}\mathbb{E}[\phi_{\mathcal{Z}}(Z)|W,A] \otimes \phi_{\mathcal{A}}(A) \otimes \phi_{\mathcal{A}}(a')\| \leq \kappa^3. \end{aligned}$$

Now, we apply Corollary (12.48) such that with probability at least $1-\delta$,

$$\|\tilde{g}_4 - g_4\| \leq 2\kappa^3 \sqrt{\frac{2\log(2/\delta)}{m}}.$$

Now, consider the decomposition

$$\begin{aligned} \|\bar{g}_{4,m} - g_4\| &\leq \|\bar{g}_{4,m} - \tilde{g}_4\| + \|\tilde{g}_4 - g_4\| \\ &\leq \|\bar{g}_{4,m} - \tilde{g}_4\| + 2\kappa^3 \sqrt{\frac{2\log(2/\delta)}{m}} \end{aligned}$$

Next, we bound the component $\|\bar{g}_{4,m} - \tilde{g}_4\|$.

$$\begin{aligned} \|\bar{g}_{4,m} - \tilde{g}_4\| &= \Bigg\|\frac{1}{m}\sum_{j=1}^m C_{Z|W,A}\left(\hat{E}[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(a_j)\right) \otimes \phi_{\mathcal{A}}(a_j) \otimes \phi_{\mathcal{A}}(a') \\ &\quad - \frac{1}{m}\sum_{j=1}^m C_{Z|W,A}\left(E[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(a_j)\right) \otimes \phi_{\mathcal{A}}(a_j) \otimes \phi_{\mathcal{A}}(a')\Bigg\| \\ &= \Bigg\|\frac{1}{m}\sum_{j=1}^m C_{Z|W,A}\left((\hat{E}[\phi_{\mathcal{W}}(W)|A=a'] - E[\phi_{\mathcal{W}}(W)|A=a']) \otimes \phi_{\mathcal{A}}(a_j)\right) \otimes \phi_{\mathcal{A}}(a_j) \otimes \phi_{\mathcal{A}}(a')\Bigg\| \\ &\leq \frac{1}{m}\sum_{j=1}^m \kappa^3 \left\|C_{Z|W,A}\right\| \left\|\hat{E}[\phi_{\mathcal{W}}(W)|A=a'] - E[\phi_{\mathcal{W}}(W)|A=a']\right\| \\ &= \kappa^3 \left\|C_{Z|W,A}\right\| \left\|\hat{E}[\phi_{\mathcal{W}}(W)|A=a'] - E[\phi_{\mathcal{W}}(W)|A=a']\right\| \end{aligned}$$

Appealing to the Theorem (12.39),

$$\left\|\hat{E}[\phi_{\mathcal{W}}(W)|A=a'] - E[\phi_{\mathcal{W}}(W)|A=a']\right\| \leq \kappa J_5 \log(4/\delta) \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}},$$

with probability at least $1-\delta$ for $\delta \in (0,1)$. As a result,

$$\|\bar{g}_{4,m} - \tilde{g}_4\| \leq \kappa^4 J_5 \log(4/\delta) \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} \left\|C_{Z|W,A}\right\|.$$

This implies that

$$\|\bar{g}_{4,m} - g_4\| \leq \kappa^4 J_5 \log(4/\delta) \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} \left\|C_{Z|W,A}\right\| + 2\kappa^3 \sqrt{\frac{2\log(2/\delta)}{m}}$$

with probability $1-2\delta$ for $\delta \in (0,1)$. □

Theorem 12.41. *Suppose Assumptions (12.1), (12.29), and (12.30) hold. Then, with probability at least* $1-2\delta$ *for all* $\delta \in (0,1)$*, for* $m \geq 8\kappa^6 \log(2/\delta) g_{\lambda_2} \lambda_2^{-1}$*,*

$$\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq J_4 \frac{\log(4/\delta)}{\sqrt{\lambda_2}} \left(\sqrt{\frac{1}{m}\left(\frac{1}{\lambda_2^{p_4}} + \frac{1}{m\lambda_2}\right)} + \frac{1}{\sqrt{\lambda_2}} \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} + \sqrt{\frac{1}{\lambda_2 m}} \right) + \lambda_2^{\frac{\beta_4 - 1}{2}},$$

where J_4 *is a constant depending on* κ, β_4, B_4, D*.*

Proof. Combining the bounds in Lemma (12.33), Lemma (12.36) and Lemma (12.40) with a union bound, we obtain that with probability at least $1-2\delta$,

$$\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq \mathring{J} \Bigg(\frac{\log(4/\delta)}{\sqrt{\lambda_2}} \Bigg(\sqrt{\frac{1}{m}\left(\mathcal{N}(\lambda_2)\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{ZAA}}} + \frac{\|\bar{\varphi}_0\|^2_{\mathcal{H}_{\mathcal{ZAA}}}}{m\lambda_2}\right)} + \frac{1}{\sqrt{\lambda_2}} \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} \|C_{Z|W,A}\|$$
$$+ \sqrt{\frac{1}{\lambda_2 m}} \Bigg) + \lambda_2^{\frac{\beta_4 - 1}{2}} \Bigg),$$

where $\mathring{J}$ is a constant depending on κ, B_4. Under Assumption (12.30), using Proposition (12.35), there is a constant $D > 0$ such that $\mathcal{N}(\lambda_2) \leq D\lambda_2^{-p_4}$. Furthermore, under Assumption (12.29),

$$\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} = \|\Sigma_4^{\frac{\beta_4 - 1}{2}} \Sigma_4^{-\frac{\beta_4 - 1}{2}} \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq \kappa^{\frac{\beta_4 - 1}{2}} B_4.$$

□

Next, we will derive a bound for $\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|$.

Lemma 12.42. *Under the assumptions of Theorems (12.11) and (12.39), with probability at least* $1-2\delta$ *for* $\delta \in (0,1)$*, the following bound holds*

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\| \leq \frac{1}{\lambda_2}\kappa^3 r_1(\delta, n, \beta_1, p_1) \left(\kappa J_5 \log(4/\delta) \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} + \kappa \right)$$
$$+ \frac{1}{\lambda_2}\left(\kappa^8 r_1(\delta, n, \beta_1, p_1)^2 + 2B_1 \kappa^{8 + \frac{\beta_1 - 1}{2}} r_1(\delta, n, \beta_1, p_1)\right) \|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZA}}}.$$

Proof.

$$\begin{aligned}
&\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} = \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{g}_{4,m} - \bar{g}_{4,m} + \bar{g}_{4,m}) - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&= \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{g}_{4,m} - \bar{g}_{4,m}) + (\hat{\Sigma}_{2,m} + \lambda_2 I)^{-1}\bar{g}_{4,m} - \bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&= \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{g}_{4,m} - \bar{g}_{4,m}) + (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}\bar{g}_{4,m} - (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{4,m} + \lambda_2 I)\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&= \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{g}_{4,m} - \bar{g}_{4,m}) + (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}\bar{g}_{4,m} - (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{4,m}\bar{\varphi}_{\lambda_4,m} + \underbrace{\lambda_2 \bar{\varphi}_{\lambda_2,m}}_{\bar{g}_{4,m} - \bar{\Sigma}_{4,m}\bar{\varphi}_{\lambda_2,m}})\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&= \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{g}_{4,m} - \bar{g}_{4,m}) + (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}\bar{g}_{4,m} - (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}\bar{g}_{4,m} \\
&- (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{4,m}\bar{\varphi}_{\lambda_2,m} - \bar{\Sigma}_{4,m}\bar{\varphi}_{\lambda_2,m})\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&= \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{g}_{4,m} - \bar{g}_{4,m}) - (\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}(\hat{\Sigma}_{4,m} - \bar{\Sigma}_{4,m})\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&\leq \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}\|_{op}\|\hat{g}_{4,m} - \bar{g}_{4,m}\|_{\mathcal{H}_{\mathcal{ZA}}} + \|(\hat{\Sigma}_{4,m} + \lambda_2 I)^{-1}\|_{op}\|\hat{\Sigma}_{4,m} - \bar{\Sigma}_{4,m}\|_{op}\|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} \\
&\leq \lambda_2^{-1}\left(\|\hat{g}_{4,m} - \bar{g}_{4,m}\|_{\mathcal{H}_{\mathcal{ZAA}}} + \|\hat{\Sigma}_{4,m} - \bar{\Sigma}_{4,m}\|_{op}\|\bar{\varphi}_{\lambda_2,m}\|_{\mathcal{H}_{\mathcal{ZAA}}}\right).
\end{aligned}$$

We have two terms to bound. First, we observe that

$$\|\hat{g}_{4,m} - \bar{g}_{4,m}\| = \left\| \frac{1}{m} \sum_{\substack{i,j=1 \\ j \neq i}}^{m} \theta_i \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a') - \frac{1}{m} \sum_{\substack{i,j=1 \\ j \neq i}}^{m} \theta_i \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a') \right\|$$

$$
\begin{aligned}
&\leq \kappa^2 \left\| \frac{1}{m} \sum_{\substack{i,j=1\\ j\neq i}}^{m} \theta_i \hat{C}_{Z|W,A}(\phi_{\mathcal{W}}(\tilde{w}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_j)) - \frac{1}{m} \sum_{\substack{i,j=1\\ j\neq i}}^{m} \theta_i C_{Z|W,A}(\phi_{\mathcal{W}}(\tilde{w}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_j)) \right\| \\
&= \kappa^2 \left\| \frac{1}{m} \sum_{j=1}^{m} \hat{C}_{Z|W,A}\Big(\sum_{\substack{i=1\\ i\neq j}}^{m} (\theta_i \phi_{\mathcal{W}}(\tilde{w}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_j)\Big) - \frac{1}{m} \sum_{j=1}^{m} C_{Z|W,A}\Big(\sum_{\substack{i=1\\ i\neq j}}^{m} (\theta_i \phi_{\mathcal{W}}(\tilde{w}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_j)\Big) \right\| \\
&= \kappa^2 \left\| \frac{1}{m} \sum_{j=1}^{m} \hat{C}_{Z|W,A}\Big(\hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(\tilde{a}_j)\Big) - \frac{1}{m} \sum_{j=1}^{m} C_{Z|W,A}\Big(\hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] \otimes \phi_{\mathcal{A}}(\tilde{a}_j)\Big) \right\| \\
&\leq \frac{\kappa^2}{m} \sum_{j=1}^{m} \|\hat{C}_{Z|W,A} - C_{Z|W,A}\| \|\hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a']\| \|\phi_{\mathcal{A}}(a_j)\| \\
&\leq \kappa^3 r_1(\delta, n, \beta_1, p_1) \left\| \hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] - \mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] + \mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \right\| \\
&\leq \kappa^3 r_1(\delta, n, \beta_1, p_1) \Big(\left\| \hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] - \mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \right\| + \left\| \mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \right\| \Big) \\
&\leq \kappa^3 r_1(\delta, n, \beta_1, p_1) \Big(\left\| \hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] - \mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] \right\| + \kappa \Big) \\
&\quad \text{(with probability } 1-\delta\text{, Corollary (12.39))} \\
&\leq \kappa^3 r_1(\delta, n, \beta_1, p_1) \left(\kappa J_5 \log(4/\delta) \left(\frac{1}{\sqrt{m}} \right)^{\frac{\beta_5 - 1}{\beta_5 + p_3}} + \kappa \right) \quad \text{(with probability } 1-2\delta\text{, Theorem (12.39).}
\end{aligned}
$$

For the second component, we note that for $i = 1, \dots, m$,

$$
\begin{aligned}
\xi_i &:= (\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \otimes (\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \\
&\quad - (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \otimes (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \\
&= ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \otimes ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \\
&\quad + ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \otimes (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \\
&\quad + (\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')) \otimes ((\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_i)) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')),
\end{aligned}
$$

and therefore, under Assumption (12.1), by the triangular inequality and by Corollary (12.12)),

$$
\begin{aligned}
\|\xi_i\|_{op} &\leq \kappa^4 \left\| \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}}^2 + 2\kappa^4 \left\| \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} \left\| \hat{\mu}_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) - \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} \\
&\leq \kappa^8 r_1(\delta, n, \beta_1, p_1)^2 + 2\kappa^6 \left\| \mu_{Z|W,A}(\tilde{w}_j, \tilde{a}_i) \right\|_{\mathcal{H}_{\mathcal{Z}}} r_1(\delta, n, \beta_1, p_1).
\end{aligned}
$$

with probability at least $1-\delta$. Also, recall that we observe in the proof of Lemma (12.20) that

$$
\|\mu_{Z|W,A}(w, a)\|_{\mathcal{H}_{\mathcal{Z}}} \leq B_1 \kappa^{2 + \frac{\beta_1 - 1}{2}}.
$$

As a result, we have the following bound with probability at least $1-\delta$,

$$
\begin{aligned}
\|\hat{\Sigma}_{4,m} - \bar{\Sigma}_{4,m}\|_{op} = \left\| \frac{1}{m} \sum_{i=1}^{m} \xi_i \right\|_{op} &\leq \frac{1}{m} \sum_{i=1}^{m} \|\xi_i\|_{op} \\
&\leq \kappa^8 r_1(\delta, n, \beta_1, p_1)^2 + 2B_1 \kappa^{8 + \frac{\beta_1 - 1}{2}} r_1(\delta, n, \beta_1, p_1). \qquad (33)
\end{aligned}
$$

□

The following Theorem provides convergence rates in RKHS norm for the estimation of the bridge solution with minimum RKHS norm.

Theorem 12.43. *Suppose Assumptions (5.1-1), (12.1), (12.6-1), (12.9-1), (12.25),(12.29), (12.30), (12.37). (12.38) hold and set* $\lambda_1 = \Theta\left(n^{-\frac{1}{\beta_1 + p_1}}\right)$, $\zeta = \Theta\left(m^{-\frac{1}{\beta_5 + p_3}}\right)$ *and* $n = m^{\iota \frac{(\beta_4 + 1)(\beta_1 + p_1)}{(\beta_4 + p_4)(\beta_1 - 1)}}$ *where* $\iota > 0$. *Then,* *i.* *If* $\iota \leq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$ *then* $\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} = O_p\left(m^{-\frac{\iota(\beta_4-1)}{2(\beta_4+p_4)}}\right)$ *with* $\lambda_2 = \Theta\left(m^{\frac{-\iota}{\beta_4+p_4}}\right)$,

ii. *If* $\iota \geq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$ *then* $\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} = O_p\left(m^{-\frac{(\beta_4-1)(\beta_5-1)}{2(\beta_4+1)(\beta_5+p_3)}}\right)$ *with* $\lambda_2 = \Theta\left(m^{\frac{-(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)}}\right)$.

Proof. Let us abbreviate $r_1(n) = r_1(\delta, n, \beta_1, p_1)$. From Lemma (12.42), we obtain with high probability that

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} \leq \|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} + \frac{\bar{J}}{\lambda_2} r_1(n)\left(1 + \left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5-1}{\beta_5+p_3}}\right)$$
$$+ \frac{\bar{J}}{\lambda_2}(1 + r_1(n)) r_1(n) (\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}} + \|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}}}),$$

where $\bar{J}$ is a constant depending on $\kappa, B_1, \beta_1, J_5$. Furthermore, from Theorem (12.41),

$$\|\bar{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} = O_p\left(r_2(m)\right),$$

with

$$r_2(m) = \frac{1}{\sqrt{\lambda_2}}\left(\sqrt{\frac{1}{m}\left(\frac{1}{\lambda_2^{p_4}} + \frac{1}{m\lambda_2}\right)} + \frac{1}{\sqrt{\lambda_2}}\left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5-1}{\beta_5+p_3}} + \sqrt{\frac{1}{\lambda_2 m}}\right) + \lambda_2^{\frac{\beta_4-1}{2}}$$
$$= \sqrt{\frac{1}{m\lambda^{p_4+1}}\left(1 + \frac{1}{m\lambda_2^{1-p_4}}\right)} + \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5-1}{\beta_5+p_3}} + \frac{1}{\lambda_2\sqrt{m}} + \lambda_2^{\frac{\beta_4-1}{2}}$$

as long as $m \geq 8\kappa^6 \log(2/\delta) g_{\lambda_2}\lambda_2$. Note that the term $\frac{1}{\lambda_2\sqrt{m}}$ is of faster order with respect to $\frac{1}{\lambda_2}\left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5-1}{\beta_5+p_3}}$ and can be removed. Discarding the other faster terms similar to Theorem (12.21) and putting it together, we obtain

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{Z}\mathcal{A}\mathcal{A}}} = O_p\left(\sqrt{\frac{1}{m\lambda^{p_4+1}}\left(1 + \frac{1}{m\lambda_2^{1-p_4}}\right)} + \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{m}}\right)^{\frac{\beta_5-1}{\beta_5+p_3}} + \lambda_2^{\frac{\beta_4-1}{2}} + \frac{1}{\lambda_2}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\beta_1-1}{\beta_1+p_1}}\right)$$

In this proof, similar to Theorem (12.21) we check, $\lambda_2^{-1}m^{-1}\log\lambda_2^{-1} = O(1)$.

Case i. Let $\iota \leq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$ and $\lambda_2 = m^{\frac{-\iota}{\beta_4+p_4}}$. We need to check that $\lambda_2^{-1}m^{-1}\log\lambda_2^{-1} = O(1)$.

$$\frac{\log\lambda_2^{-1}}{m\lambda_2} = \frac{\iota}{\beta_4+p_4}\frac{\log m}{m^{1-\frac{\iota}{\beta_4+p_4}}}.$$

As $\frac{\iota}{\beta_4+p_4} \leq \frac{\beta_5-1}{(\beta_5+p_3)(\beta_4+1)} \leq 1$, we have $\lambda_2^{-1}m^{-1}\log\lambda_2^{-1} \to 0$ as $m \to \infty$. Next, note that with this choice of λ_2 we have

$$\lambda_2^{\frac{\beta_4-1}{2}} = m^{-\frac{\iota}{2}\frac{\beta_4+1}{\beta_4+p_4}} = \frac{1}{\lambda_2}n^{-\frac{1}{2}\frac{\beta_1-1}{\beta_1+p_1}}.$$

Furthermore,

i.

$$\lambda_2^{\beta_4-1} \geq \frac{1}{m\lambda_2^{p_4+1}} \iff \lambda_2^{\beta_4+p_4} \geq m^{-1}$$
$$\iff m^{\iota\frac{\beta_4+p_4}{\beta_4+p_4}} \leq m \iff \iota \leq 1$$

that is true due to our assumption in this condition $\iota \leq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)} \leq 1$.

ii.

$$\lambda_2^{\frac{\beta_4-1}{2}} \geq \frac{1}{\lambda_2} m^{-\frac{1}{2}\frac{\beta_5-1}{\beta_5+p_3}} \iff \lambda_2^{\frac{\beta_4+1}{2}} \geq m^{-\frac{1}{2}\frac{\beta_5-1}{\beta_5+p_3}}$$
$$\iff m^{-\frac{\iota}{2}\frac{\beta_4+1}{\beta_4+p_4}} \geq m^{-\frac{1}{2}\frac{\beta_5-1}{\beta_5+p_3}}$$
$$\iff \iota \leq \frac{(\beta_4+p_4)(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)}$$

iii.

$$\frac{1}{m\lambda_2^{1-p_4}} = \frac{1}{mm^{-\frac{\iota(1-p_4)}{\beta_4+p_4}}} = \frac{1}{m^{1-\frac{\iota(1-p_4)}{\beta_4+p_4}}} \leq 1$$

since $\beta_4 + p_4 - \iota(1-p_4) \geq 0$ due to the fact that $\iota \leq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)} \leq 1$.

iv. Also

$$\frac{1}{m\lambda_2^{1-p_4}} \leq \frac{1}{m\lambda^{1+p_4}}.$$

Therefore

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} = O\left(\lambda_2^{\frac{\beta_4-1}{2}}\right) = O\left(m^{-\frac{\iota}{2}\frac{\beta_4-1}{\beta_4+p_4}}\right).$$

Case ii. Let $\iota \geq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$ and $\lambda_2 = m^{\frac{-(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)}}$. Let us first check $\lambda_2^{-1}m^{-1}\log(\lambda_2^{-1}) = O(1)$.

$$\frac{\log\lambda_2^{-1}}{m\lambda_2} = \frac{\beta_5-1}{(\beta_5+p_3)(\beta_4+1)}\frac{\log m}{m^{1-\frac{\beta_5-1}{(\beta_5+p_3)(\beta_4+1)}}}.$$

As $\frac{\beta_5-1}{(\beta_5+p_3)(\beta_4+1)} < 1$, we have $\lambda_2^{-1}m^{-1}\log\lambda_2^{-1} \to 0$ as $m \to \infty$. Next, note that with this choice of λ_2 we have

$$\lambda_2^{\frac{\beta_4-1}{2}} = \frac{1}{\lambda_2} m^{-\frac{1}{2}\frac{\beta_5-1}{\beta_5+p_3}}.$$

Furthermore,

i.

$$\lambda_2^{\beta_4-1} \geq \frac{1}{m\lambda_2^{p_4+1}} \iff \lambda_2^{\beta_4+p_4} \geq m^{-1}$$
$$\iff m^{-\frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}} \geq m^{-1}$$

which is true since $\frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)} \leq 1$

ii.

$$\lambda_2^{\frac{\beta_4-1}{2}} \geq \frac{1}{\lambda_2} n^{-\frac{1}{2}\frac{\beta_1-1}{\beta_1+p_1}} \iff \lambda_2^{\frac{\beta_4+1}{2}} \geq n^{-\frac{1}{2}\frac{\beta_1-1}{\beta_1+p_1}}$$
$$\iff m^{-\frac{1}{2}\frac{(\beta_5-1)(\beta_4+1)}{(\beta_5+p_3)(\beta_4+1)}} \geq m^{-\frac{\iota}{2}\frac{(\beta_1-1)(\beta_4+1)(\beta_1+p_1)}{(\beta_1+p_1)(\beta_4+p_4)(\beta_1-1)}}$$

which is true since $\iota \geq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$.

iii.

$$\frac{1}{m\lambda_2^{1-p_4}} = \frac{1}{mm^{\frac{-(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)}}} = \frac{1}{m^{1-\frac{(1-p_4)(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)}}} \leq 1$$

since $\frac{(1-p_4)(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)} \leq 1$.

iv. Also,

$$\frac{1}{m\lambda_2^{1-p_4}} \leq \frac{1}{m\lambda_2^{1+p_4}}.$$

Hence,

$$\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} = O\left(\lambda_2^{\frac{\beta_4-1}{2}}\right) = O\left(m^{-\frac{1}{2}\frac{(\beta_5-1)(\beta_4-1)}{(\beta_4+1)(\beta_5+p_3)}}\right).$$

□

12.2.2 Consistency for Conditional Dose-Response Curve

Theorem 12.44. *Suppose that Assumptions (5.1), (12.1), (12.6-1 & 3) and (12.9-1 & 3), (12.29), (12.30) hold and set* $\lambda_1 = \Theta\left(n^{-\frac{1}{\beta_1+p_1}}\right)$, $\lambda_3 = \Theta\left(t^{-\frac{1}{\beta_3+p_3}}\right)$, $\zeta = \Theta\left(m^{-\frac{1}{\beta_5+p_3}}\right)$, *and* $n = m^{\iota\frac{(\beta_4+1)(\beta_1+p_1)}{(\beta_4+p_4)(\beta_1-1)}}$ *where* $\iota > 0$. *Then,*

i. If $\iota \leq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$ *then* $\sup_a |\hat{f}_{ATT}(a,a') - f_{ATT}(a,a')| = O_p\left(\sqrt{t}^{-\frac{\beta_3-1}{\beta_3+p_3}} + m^{-\frac{\iota(\beta_4-1)}{2(\beta_4+p_4)}}\right)$ *with* $\lambda_2 = \Theta\left(m^{\frac{-\iota}{\beta_4+p_4}}\right)$,

ii. If $\iota \geq \frac{(\beta_5-1)(\beta_4+p_4)}{(\beta_5+p_3)(\beta_4+1)}$ *then* $\sup_a |\hat{f}_{ATT}(a,a') - f_{ATT}(a,a')| = O_p\left(\sqrt{t}^{-\frac{\beta_3-1}{\beta_3+p_3}} + m^{-\frac{(\beta_4-1)(\beta_5-1)}{2(\beta_4+1)(\beta_5+p_3)}}\right)$ *with* $\lambda_2 = \Theta\left(m^{\frac{-(\beta_5-1)}{(\beta_4+1)(\beta_5+p_3)}}\right)$.

Proof. For fix a' and for any $a \in \mathcal{A}$, we apply the following decomposition,

$$\begin{aligned}
&|\hat{f}_{ATT}(a,a') - f_{ATT}(a,a')| = |\langle \hat{\varphi}_{\lambda_2,m}, \hat{\Psi}(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}} - \langle \bar{\varphi}_0, \Psi(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}}| \\
&= |\langle \hat{\varphi}_{\lambda_2,m}, (\hat{\Psi} - \Psi)(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}} + \langle (\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0), \Psi(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}}| \\
&= |\langle \hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0, (\hat{\Psi} - \Psi)(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}} + \langle \bar{\varphi}_0, (\hat{\Psi} - \Psi)(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}} \qquad (34) \\
&+ \langle (\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0), \Psi(a) \otimes \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{A}}(a')\rangle_{\mathcal{H}_{\mathcal{ZAA}}}| \\
&\leq \kappa^2 \left(\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} + \|\bar{\varphi}_0\| \|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{ZA}}} + \|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \|\Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}}\right) \qquad (35)
\end{aligned}$$

Note that, under Assumption (5.1-3), Assumption (12.1) and Assumption (12.6-3),

$$\|\Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} = \|C_{YZ|A}\phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{Z}}} \leq \kappa \|C_{YZ|A}\Sigma_3^{-\frac{\beta_3-1}{2}}\Sigma_3^{\frac{\beta_3-1}{2}}\|_{S_2(\mathcal{H}_{\mathcal{A}},\mathcal{H}_{\mathcal{Z}})} \leq B_3\kappa^{1+\frac{\beta_3-1}{2}} =: \alpha_1.$$

Furthermore, under Assumption (12.29), by Proposition (12.32),

$$\|\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} = \|\Sigma_2^{\frac{\beta_4-1}{2}}\Sigma_2^{-\frac{\beta_4-1}{2}}\bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZAA}}} \leq B_2\kappa^{\frac{\beta_4-1}{2}} =: \alpha_2.$$

Therefore,

$$|\hat{f}_{ATT}(a,a') - f_{ATT}(a,a')| \leq \kappa^2 \left(\|\hat{\varphi}_{\lambda_2,m} - \bar{\varphi}_0\|_{\mathcal{H}_{\mathcal{ZA}}} \|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{Z}}} + \alpha_2 \|\hat{\Psi}(a) - \Psi(a)\|_{\mathcal{H}_{\mathcal{ZA}}} + \alpha_1 \|\hat{\varphi}_{\lambda_2,m} - \varphi_0\|_{\mathcal{H}_{\mathcal{ZA}}}\right).$$

As the first term is faster than the two last terms, plugging the results of Theorem (12.43) and Corollary (12.23), we obtain the final bound. □

12.3 Additional results

Here, we present the results that we used in deriving the consistency bounds for dose-response and conditional dose-response estimations.

The following is a direct consequence of Fischer and Steinwart (2020, Lemma 17).

Lemma 12.45. *For any $\delta \in (0,1)$ and $m \geq 8\kappa^4 \log(2/\delta) g_{\lambda_2} \lambda_2^{-1}$ where $g_{\lambda_2} = \log\left(2e\mathcal{N}(\lambda_2)\frac{\|\Sigma_2\|_{op}+\lambda_2}{\|\Sigma_2\|_{op}}\right)$ and under Assumption 12.1*

$$\left\|(\Sigma_2 + \lambda_2 I)^{1/2}(\bar{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\Sigma_2 + \lambda_2 I)^{1/2}\right\|_{op} \leq 3,$$

Proof. By Fischer and Steinwart (2020, Lemma 17), for $\delta \in (0,1), \lambda_2 > 0$, and $m \geq 1$, the following operator norm bound is satisfied probability not less than $1-\delta$,

$$\left\|(\Sigma_2 + \lambda_2 I)^{-1/2}\left(\Sigma_2 - \bar{\Sigma}_{2,m}\right)(\Sigma_2 + \lambda_2 I)^{-1/2}\right\| \leq \frac{4\kappa^4 \log(2/\delta) g_{\lambda_2}}{3m\lambda_2} + \sqrt{\frac{2\kappa^4 \log(2/\delta) g_{\lambda_2}}{m\lambda_2}},$$

where we took $\alpha = 1$ in their result since the kernels are bounded. In their notations, for $\alpha = 1$ (see Fischer and Steinwart (2020, Eq. (15) & Eq. (16))

$$\|k_\nu^\alpha\|_\infty = \sup_{(w,a)\in\mathcal{W}\times\mathcal{A}} \|\mu_{Z|W,A}(w,a) \otimes \phi_{\mathcal{A}}(a)\|_{\mathcal{H}_{\mathcal{WA}}} \leq \kappa^2.$$

Therefore, with $m \geq 8\kappa^4 \log(2/\delta) g_{\lambda_2} \lambda_2^{-1}$,

$$\left\|(\Sigma_2 + \lambda_2 I)^{-1/2}\left(\Sigma_2 - \bar{\Sigma}_{2,m}\right)(\Sigma_2 + \lambda_2 I)^{-1/2}\right\| \leq \frac{4}{3}\cdot\frac{1}{8} + \sqrt{2\cdot\frac{1}{8}} = \frac{2}{3}$$

with probability not less than $1-\delta$. Consequently, the inverse of

$$\mathrm{Id} - (\Sigma_2 + \lambda_2 I)^{-1/2}\left(\Sigma_2 - \bar{\Sigma}_{2,m}\right)(\Sigma_2 + \lambda_2 I)^{-1/2}$$

can be represented by the Neumann series. In particular, the Neumann series gives us the following bound,

$$\begin{aligned}
&\left\|(\Sigma_2 + \lambda_2 I)^{1/2}(\bar{\Sigma}_{2,m} + \lambda_2 I)^{-1}(\Sigma_2 + \lambda_2 I)^{1/2}\right\|^2 \\
&= \left\|\mathrm{Id} - (\Sigma_2 + \lambda_2 I)^{-1/2}\left(\Sigma_2 - \bar{\Sigma}_{2,m}\right)(\Sigma_2 + \lambda_2 I)^{-1/2}\right\|^2 \\
&\leq \left(\sum_{k=0}^{\infty}\left\|(\Sigma_2 + \lambda_2 I)^{-1/2}\left(\Sigma_2 - \bar{\Sigma}_{2,m}\right)(\Sigma_2 + \lambda_2 I)^{-1/2}\right\|^k\right)^2 \\
&\leq \left(\sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k\right)^2 = 9
\end{aligned}$$

□

Lemma 12.46 (Lemma 25 Fischer and Steinwart (2020))**.** *Let, for $\lambda > 0$ and $0 \leq \alpha \leq 1$, the function $f_{\lambda,\alpha} : [0,\infty) \to \mathbb{R}$ be defined by $f_{\lambda,\alpha}(t) := t^\alpha/(\lambda + t)$. The supremum of $f_{\lambda,\alpha}$ satisfies the following bound*

$$\sup_{t\geq 0} f_{\lambda,\alpha}(t) \leq \lambda^{\alpha-1}.$$

The following result is the direct consequence of (Pinelis, 1994, Theorem 3.5).

Proposition 12.47 (Hoeffing's inequality in Hilbert space)**.** *Let $\xi_1, \ldots, \xi_m$ be independent centered random variables taking values in a separable Hilbert space $\mathcal{H}$ such that $\|\xi_i\|_{\mathcal{H}} \leq M$ almost surely for all $1 \leq i \leq m$. Then for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have,*

$$\left\|\frac{1}{m}\sum_{i=1}^{n}\xi_i\right\|_{\mathcal{H}} \leq M\sqrt{\frac{2\log(2/\delta)}{m}}$$

We obtain the following inequality for non-centered random variables.

Corollary 12.48 (Hoeffing's inequality in Hilbert space)**.** *Let* $\xi_1, \dots, \xi_m$ *be independent (not necessarily centered) random variables taking values in a separable Hilbert space* $\mathcal{H}$ *such that* $\|\xi_i\|_{\mathcal{H}} \le M$ *almost surely for all* $1 \le i \le m$*. Then for all* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, we have,*

$$\left\| \frac{1}{m} \left(\sum_{i=1}^{n} \xi_i - \mathbb{E}[\xi_i] \right) \right\|_{\mathcal{H}} \le 2M \sqrt{\frac{2\log(2/\delta)}{m}}$$

Proof. By Jensen's inequality and the triangular inequality, for all $1 \le i \le m$,

$$\|\xi_i - \mathbb{E}[\xi_i]\|_{\mathcal{H}} \le \|\xi_i\|_{\mathcal{H}} + \|\mathbb{E}[\xi_i]\|_{\mathcal{H}} \le \|\xi_i\|_{\mathcal{H}} + \mathbb{E}[\|\xi_i\|_{\mathcal{H}}] \le 2M,$$

and the result follows from Proposition (12.47). □

The following Bernstein's inequality can be found in Fischer and Steinwart (2020, Theorem 26).

Theorem 12.49 (Bernstein's inequality in Hilbert space)**.** *Let* $\xi_1, \dots, \xi_m$ *be independent and identically distributed random variables taking values in a separable Hilbert space* $\mathcal{H}$ *such that*

$$\mathbb{E}[\|\xi_1\|_H^q] \le \frac{1}{2} q! \sigma^2 L^{q-2}$$

for all $q \ge 2$*. Then for all* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, we have,*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \xi_i - \mathbb{E}\xi_i \right\|_H \le \log(2/\delta) \sqrt{\frac{32}{m} \left(\sigma^2 + \frac{L^2}{m} \right)}.$$

13 SUPPLEMENTARY ON NUMERICAL EXPERIMENTS

In this section, we provide more detail on the numerical experiments as well as further ablation studies. We first provide details on the kernel function that we used in our experiments and the procedure to tune the regularization parameters. After that, we provide further information about the experimental setups and discuss the complexity of our proposed methods.

13.1 Kernel

In our experiments, we employed the Gaussian kernel function

$$k_{\mathcal{F}}(f_i, f_j) = \exp\left(\frac{-\|f_i - f_j\|_2^2}{2l^2} \right) \tag{36}$$

for $f_i, f_j \in \mathbb{R}^{d_{\mathcal{F}}}$. Gaussian kernel is bounded, continuous and characteristic. The parameter l is called the length scale of the kernel and there is a simple heuristic to determine the length scale called *median length scale heuristic*. In particular, consider the data $\{f_i\}_{i=1}^n$. Then, we set the length scale squared l^2 to the half of the median value of the pairwise squared distances, i.e.,

$$l^2 = \frac{1}{2} \text{median}(\{\|f_i - f_j\|_2^2 : 1 \le i < j \le n\}).$$

This heuristic has also been utilized in causal inference literature, e.g., (Singh et al., 2023; Mastouri et al., 2021; Singh, 2023; Xu and Gretton, 2024).

The Gaussian kernel in Equation (36) can also be considered as multiplication of Gaussian kernels for each dimension, i.e.,

$$k_{\mathcal{F}}(f_i, f_j) = \prod_{k=1}^{d_{\mathcal{F}}} \exp\left(\frac{-\|f_i^{(k)} - f_j^{(k)}\|_2^2}{2{l^{(k)}}^2} \right) \tag{37}$$

where $f_i^{(k)}$ is the k-th dimension of $f_i \in \mathbb{R}^{d_{\mathcal{F}}}$. In that case, each of the length scale in the set $\{l^{(k)}\}_{k=1}^{d_{\mathcal{F}}}$ can be determined by the median distance length-scale heuristic for each dimension separately. We will refer to the kernel in Equation (37) as *columnwise Gaussian kernel.* In our synthetic low-dimensional experiment in Section (6.1) we used columnwise Gaussian kernel for outcome proxy variable W. For all the other experiments and variables with our proposed method, we used Gaussian kernel.

13.2 Hyperparameter Selection

13.2.1 Tuning λ_1 and λ_2 Regularization Parameters

To tune the regularization parameters λ_1 and λ_3, we employ leave-one-out cross validation (LOOCV) technique since it has a closed-form expression in the kernel ridge regression setup. The following theorem provides the closed-form expression for the LOOCV loss in kernel ridge regression.

Theorem 13.1 (Theorem F.1 in (Singh et al., 2023))**.** *Consider the kernel ridge regression setup from measurements $\{x_i\}_{i=1}^n$ to the outcomes $\{y_i\}_{i=1}^n$, and we minimize the regularized squared loss:*

$$\begin{aligned}\arg\min_{f \in \mathcal{H}_\mathcal{X}} \mathcal{L}(f) &= \arg\min_{f \in \mathcal{H}_\mathcal{X}} \left[\frac{1}{n}\sum_{i=1}^n (y_i - \langle f, \phi_\mathcal{X}(x_i)\rangle_\mathcal{H})^2 + \lambda \|f\|_\mathcal{H}^2\right] \\ &= \arg\min_{f \in \mathcal{H}_\mathcal{X}} \left[\frac{1}{n}\|\boldsymbol{Y} - \Phi_\mathcal{X}^T f\|_2^2 + \lambda \|f\|_\mathcal{H}^2\right]\end{aligned}$$

where $\phi_\mathcal{X}(.)$ is the canonical feature map for the assumed kernel function $k_\mathcal{X}(.,.)$, $\mathcal{H}_\mathcal{X}$ is the corresponding RKHS, $\boldsymbol{Y} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T \in \mathbb{R}^n$ and $\Phi_\mathcal{X} = \begin{bmatrix} \phi_\mathcal{X}(x_1) & \phi_\mathcal{X}(x_2) & \dots & \phi_\mathcal{X}(x_n) \end{bmatrix}$. Then, the LOOCV loss is given by

$$LOOCV_f(\lambda) = \frac{1}{n}\|\tilde{\boldsymbol{H}}_\lambda^{-1}\boldsymbol{H}_\lambda \boldsymbol{Y}\|_2^2 \tag{38}$$

where

$$\begin{aligned}\boldsymbol{H}_\lambda &= \boldsymbol{I} - \boldsymbol{K}_{\boldsymbol{XX}}(\boldsymbol{K}_{\boldsymbol{XX}} + n\lambda \boldsymbol{I})^{-1} \in \mathbb{R}^{n \times n} \\ \tilde{\boldsymbol{H}}_\lambda &= diag(\boldsymbol{H}_\lambda) \in \mathbb{R}^{n \times n}\end{aligned}$$

The proof of this theorem can be found in (Singh et al., 2023) (see Algorithm F.1 in that paper). As a result o this theorem, one can tune the hyperparameter λ of kernel ridge regression over a grid $\Lambda \subset \mathbb{R}$, i.e.,

$$\lambda^* = \arg\min_{\Lambda \subset \mathbb{R}} \frac{1}{n}\|\tilde{\boldsymbol{H}}_\lambda^{-1}\boldsymbol{H}_\lambda \boldsymbol{Y}\|_2^2.$$

Tuning of First Stage Regression Regularization (λ_1): Recall that in first-stage regression, we solve the following optimization problem (see derivation of Algorithm (4.1) in S.M. (11.1):

$$\hat{\mathcal{L}}^c(C) = \frac{1}{n}\sum_{i=1}^n \|\phi_\mathcal{Z}(z_i) - C(\phi_\mathcal{W}(w_i) \otimes \phi_\mathcal{A}(a_i))\|_{\mathcal{H}_\mathcal{Z}}^2 + \lambda_1 \|C\|_{S_2(\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{Z})}^2.$$

This is a kernel ridge regression from the (infinite dimensional) measurements $\{\phi_\mathcal{W}(w_i) \otimes \phi_\mathcal{A}(a_i)\}_{i=1}^n$ to the (infinite dimensional) outcomes $\{\phi_\mathcal{Z}(z_i)\}_{i=1}^n$. Since the kernel function for the tensor product features $\phi_\mathcal{W}(w_i) \otimes \phi_\mathcal{A}(a_i)$ is $k_\mathcal{W}(w_i, .)k_\mathcal{A}(a_i, .)$, we can write

$$\begin{aligned}\boldsymbol{H}_{\lambda_1} &= \boldsymbol{I} - (\boldsymbol{K}_{\boldsymbol{WW}} \odot \boldsymbol{K}_{\boldsymbol{AA}})(\boldsymbol{K}_{\boldsymbol{WW}} \odot \boldsymbol{K}_{\boldsymbol{AA}} + n\lambda_1 \boldsymbol{I})^{-1} \in \mathbb{R}^{n \times n} \\ \tilde{\boldsymbol{H}}_{\lambda_1} &= \operatorname{diag}(\boldsymbol{H}_{\lambda_1}) \in \mathbb{R}^{n \times n}.\end{aligned}$$

Furthermore, we see that the LOOCV can be written as

$$\begin{aligned}\mathrm{LOOCV}_C(\lambda) &= \frac{1}{n}\|\tilde{\boldsymbol{H}}_{\lambda_1}^{-1}\boldsymbol{H}_{\lambda_1}\Phi_\mathcal{Z}^T\|_2^2 \\ &= \frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{H}}_{\lambda_1}^{-1}\boldsymbol{H}_{\lambda_1}\Phi_\mathcal{Z}^T\Phi_\mathcal{Z}\boldsymbol{H}_{\lambda_1}^T\tilde{\boldsymbol{H}}_{\lambda_1}^{-T}) \\ &= \frac{1}{n}\mathrm{Tr}(\tilde{\boldsymbol{H}}_{\lambda_1}^{-1}\boldsymbol{H}_{\lambda_1}\boldsymbol{K}_{\boldsymbol{ZZ}}\boldsymbol{H}_{\lambda_1}^T\tilde{\boldsymbol{H}}_{\lambda_1}^{-T}).\end{aligned} \tag{39}$$

Hence, we can tune λ_1 over a grid $\Lambda_1 \subset \mathbb{R}$ that minimizes the LOOCV loss in Equation (39). In each of our experiments, we generated the grid Λ_1 with *logspace* with maximum and minimum values of 1.0 and 10^{-7}, respectively, and we used 150 grid points. This procedure applies to both Algorithms for dose-response curve and conditional dose-response curve estimations.

Tuning of Third Stage Regression Regularization (λ_3): Recall that in order to estimate the causal functions after second-stage regression, we solve the following optimization problem (see derivation of Algorithm (4.1) in S.M. (11.1):

$$\begin{aligned}\hat{C}_{YZ|A} &= \arg\min_{C} \frac{1}{n}\sum_{i=1}^{n} \|y_i\phi_{\mathcal{Z}}(z_i) - C\phi_{\mathcal{A}}(a_i)\|^2 + \lambda_2\|C\|^2 \\ &= \arg\min_{C} \frac{1}{n}\|\Phi_{\mathcal{Z}}\mathrm{diag}(\boldsymbol{Y}) - C\Phi_{\mathcal{A}}\|^2 + \lambda_3\|C\|^2.\end{aligned}$$

Again, this is kernel ridge regression from (infinite dimensional) measurement $\{\phi_{\mathcal{A}}(a_i)\}_{i=1}^n$ to the (infinite dimensional) outcomes $y_i\phi_{\mathcal{Z}}(z_i)$. Then, we construct

$$\begin{aligned}\boldsymbol{H}_{\lambda_3} &= \boldsymbol{I} - \boldsymbol{K_{AA}}(\boldsymbol{K_{AA}} + n\lambda_3\boldsymbol{I})^{-1} \in \mathbb{R}^{n\times n} \\ \tilde{\boldsymbol{H}}_{\lambda_3} &= \mathrm{diag}(\boldsymbol{H}_{\lambda_3}) \in \mathbb{R}^{n\times n}.\end{aligned}$$

Therefore, LOOCV loss can be written as

$$\begin{aligned}\mathrm{LOOCV}_F(\lambda_3) &= \frac{1}{n}\|\tilde{\boldsymbol{H}}_{\lambda_3}^{-1}\boldsymbol{H}_{\lambda_3}(\Phi_{\mathcal{Z}}\mathrm{diag}(\boldsymbol{Y}))^T\|_2^2 \\ &= \frac{1}{n}\|\tilde{\boldsymbol{H}}_{\lambda_3}^{-1}\boldsymbol{H}_{\lambda_3}\mathrm{diag}(\boldsymbol{Y})^T\Phi_{\mathcal{Z}}^T\|_2^2 \\ &= \frac{1}{n}\mathrm{Tr}\Big(\tilde{\boldsymbol{H}}_{\lambda_3}^{-1}\boldsymbol{H}_{\lambda_3}\mathrm{diag}(\boldsymbol{Y})^T\Phi_{\mathcal{Z}}^T\Phi_{\mathcal{Z}}\mathrm{diag}(\boldsymbol{Y})\boldsymbol{H}_{\lambda_3}^T\tilde{\boldsymbol{H}}_{\lambda_3}^{-T}\Big) \\ &= \frac{1}{n}\mathrm{Tr}\Big(\tilde{\boldsymbol{H}}_{\lambda_3}^{-1}\boldsymbol{H}_{\lambda_3}(\boldsymbol{K_{ZZ}}\odot\boldsymbol{Y}\boldsymbol{Y}^T)\boldsymbol{H}_{\lambda_3}^T\tilde{\boldsymbol{H}}_{\lambda_3}^{-T}\Big) \qquad (40)\end{aligned}$$

As a result, we can tune λ_3 over a grid $\Lambda_3 \subset \mathbb{R}$ that minimizes the LOOCV loss in Equation (40). In each of our experiments, we generated the grid Λ_3 with *logspace* with maximum and minimum values of 1.0 and 10^{-7}, respectively, and we used 150 grid points. This procedure applies to both Algorithms for dose-response curve and conditional dose-response curve estimations.

13.2.2 Tuning λ_2 Regularization Parameter in ATE Algorithm

Recall that in the second-stage regression, we use the stage 2 data $\{\tilde{z}_i, \tilde{w}_i, \tilde{a}_i\}_{i=1}^n$. We can estimate the out-of-sample loss of the second-stage regression using the data from first-stage $\{z_i, w_i, a_i\}_{i=1}^n$ in order to tune the regularization parameter λ_2. Then, the out-of-sample loss can be expressed as follows:

$$\begin{aligned}\hat{\mathcal{L}}^{\mathrm{Val}}(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\langle\varphi, \hat{\mu}_{Z|W,X,A}(w_i,a_i)\otimes\phi_{\mathcal{A}}(a_i)\rangle_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}}^2 \qquad (41) \\ - 2\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\Big\langle\varphi, \hat{\mu}_{Z|W,X,A}(w_j,a_i)\otimes\phi_{\mathcal{A}}(a_i)\Big\rangle_{\mathcal{H}_{\mathcal{Z}}\otimes\mathcal{H}_{\mathcal{A}}} \qquad (42)\end{aligned}$$

where

$$\varphi = \sum_{i=1}^{m}\alpha_i\hat{\mu}_{Z|W,A}(\tilde{w}_i,\tilde{a}_i)\otimes\phi_{\mathcal{A}}(\tilde{a}_i) + \frac{\alpha_{m+1}}{m(m-1)}\sum_{j=1}^{m}\sum_{\substack{l=1\\ l\neq j}}^{m}\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_j)\otimes\phi_{\mathcal{A}}(\tilde{a}_j),$$

and the set $\{\alpha_i\}_{i=1}^{m+1}$ are the optimizer of the loss function in Equation (18). Now, let us compute the closed-form expression for out-of-loss. First, consider the following inner product:

$$\begin{aligned}&\langle\varphi, \hat{\mu}_{Z|W,A}(w_i,a_i)\otimes\phi_{\mathcal{A}}(a_i)\rangle \\ &= \sum_{l=1}^{m}\alpha_l\langle\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_l)\otimes\phi_{\mathcal{A}}(\tilde{a}_l), \hat{\mu}_{Z|W,A}(w_i,a_i)\otimes\phi_{\mathcal{A}}(a_i)\rangle\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1\\ l\neq j}}^{m} \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j), \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i)\rangle \\
&= \sum_{l=1}^{m} \alpha_l \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l), \hat{\mu}_{Z|W,A}(w_i, a_i)\rangle \langle \phi_{\mathcal{A}}(\tilde{a}_l), \phi_{\mathcal{A}}(a_i)\rangle \\
&+ \frac{\alpha_{m+1}}{m(m-1)} \sum_{j=1}^{m} \sum_{\substack{l=1\\ l\neq j}}^{m} \langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j), \hat{\mu}_{Z|W,A}(w_i, a_i)\rangle \langle \phi_{\mathcal{A}}(\tilde{a}_j), \phi_{\mathcal{A}}(a_i)\rangle \\
&= \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^\top \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(w_i, a_i) k_{\mathcal{A}}(\tilde{a}_l, a_i) \\
&+ \frac{\alpha_{m+1}}{m} \sum_{j=1}^{m} \left(\sum_{\substack{l=1\\ l\neq j}}^{m} \frac{1}{m-1} \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j) \right)^\top \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(w_i, a_i) k_{\mathcal{A}}(\tilde{a}_j, a_i) \\
&= \left[\left(\boldsymbol{C}^\top \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}} \right) \alpha_{1:m} + \alpha_{m+1} \left(\boldsymbol{C}^\top \boldsymbol{K}_{ZZ} \bar{B} \odot \boldsymbol{K}_{A\tilde{A}} \right) \frac{\mathbf{1}}{m} \right]_i
\end{aligned}
$$

where

$$
\boldsymbol{C} = \left(\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I} \right)^{-1} (\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{XX} \odot \boldsymbol{K}_{AA}).
$$

As a result,

$$
\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \langle \varphi, \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i)\rangle^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{n} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}^T \boldsymbol{K}_{ZZ} \boldsymbol{C} \odot \boldsymbol{K}_{\tilde{A}A} \\ (\frac{\mathbf{1}}{m})^T \big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ} \boldsymbol{C} \odot \boldsymbol{K}_{\tilde{A}A} \big) \end{bmatrix} \begin{bmatrix} \boldsymbol{C}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}} & (\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \bar{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}) \frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}
\end{aligned}
$$

Secondly, consider the following sum of inner products:

$$
\begin{aligned}
&\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \Big\langle \varphi, \hat{\mu}_{Z|W,A}(w_j, a_i) \otimes \phi_{\mathcal{A}}(a_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \sum_{l=1}^{m} \alpha_l \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l), \hat{\mu}_{Z|W,A}(w_j, a_i) \otimes \phi_{\mathcal{A}}(a_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&+ \frac{\alpha_{m+1}}{mn(m-1)(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \sum_{r=1}^{m} \sum_{\substack{s=1\\ l\neq r}}^{m} \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_s, \tilde{a}_r) \otimes \phi_{\mathcal{A}}(\tilde{a}_r), \hat{\mu}_{Z|W,A}(w_j, a_i) \otimes \phi_{\mathcal{A}}(a_i) \Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^\top \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(w_j, a_i) k_{\mathcal{A}}(\tilde{a}_l, a_i) \\
&+ \frac{\alpha_{m+1}}{mn(m-1)(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1\\ j\neq i}}^{n} \sum_{r=1}^{m} \sum_{\substack{s=1\\ l\neq r}}^{m} \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r)^\top \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(w_j, a_i) k_{\mathcal{A}}(\tilde{a}_r, a_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{m} \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^\top \boldsymbol{K}_{ZZ} \left(\frac{1}{n-1} \sum_{\substack{j=1\\ j\neq i}}^{n} \boldsymbol{\beta}(w_j, a_i) \right) k_{\mathcal{A}}(\tilde{a}_l, a_i) \\
&+ \frac{\alpha_{m+1}}{mn} \sum_{i=1}^{n} \sum_{r=1}^{m} \left(\frac{1}{m-1} \sum_{\substack{s=1\\ l\neq r}}^{m} \boldsymbol{\beta}(\tilde{w}_s, \tilde{a}_r) \right)^\top \boldsymbol{K}_{ZZ} \left(\sum_{\substack{j=1\\ j\neq i}}^{n} \boldsymbol{\beta}(w_j, a_i) \right) k_{\mathcal{A}}(\tilde{a}_r, a_i)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\alpha_{1:m}^T\big(\boldsymbol{B}^\top \boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\mathbf{1} + \alpha_{m+1}\frac{1}{nm}\mathbf{1}^\top\big(\bar{\boldsymbol{B}}^\top \boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\mathbf{1} \\
&= \begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix}^\top \begin{bmatrix}\big(\boldsymbol{B}^\top \boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\frac{\mathbf{1}}{n} \\ (\frac{\mathbf{1}}{m})^\top\big(\bar{\boldsymbol{B}}^\top \boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\frac{\mathbf{1}}{n}\end{bmatrix}
\end{aligned}
$$

where $\bar{\boldsymbol{C}}$ is the matrix whose j-th column is given by

$$
\bar{\boldsymbol{C}} = \frac{1}{n}\sum_{\substack{l=1 \\ l\neq j}}^{n}\Big(\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{XX} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I}\Big)^{-1}(\boldsymbol{K}_{Ww_l} \odot \boldsymbol{K}_{Xx_l} \odot \boldsymbol{K}_{Aa_j}).
$$

As a result, we can write the hold-out sample loss as:

$$
\begin{aligned}
\hat{\mathcal{L}}^{\text{Val}}(\varphi) &= \frac{1}{n}\sum_{i=1}^{n}\langle \varphi, \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i)\rangle^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&\quad - 2\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\Big\langle \varphi, \hat{\mu}_{Z|W,A}(w_j, a_i) \otimes \phi_{\mathcal{A}}(a_i)\Big\rangle_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \\
&= \frac{1}{n}\begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}^T \boldsymbol{K}_{ZZ}\boldsymbol{C} \odot \boldsymbol{K}_{\tilde{A}A} \\ (\frac{\mathbf{1}}{m})^T\big(\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ}\boldsymbol{C} \odot \boldsymbol{K}_{\tilde{A}A}\big)\end{bmatrix}\begin{bmatrix}\boldsymbol{C}^T \boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}} & \big(\boldsymbol{C}^T \boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}\big)\frac{\mathbf{1}}{m}\end{bmatrix}\begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix} \\
&\quad - 2\begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix}^\top \begin{bmatrix}\big(\boldsymbol{B}^\top \boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\frac{\mathbf{1}}{n} \\ (\frac{\mathbf{1}}{m})^\top\big(\bar{\boldsymbol{B}}^\top \boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\frac{\mathbf{1}}{n}\end{bmatrix} \qquad (43)
\end{aligned}
$$

One can choose the regularization parameter λ_2 that will minimize $\mathcal{L}^{\hat{\text{Val}}}$ in Equation (43). However, even though validation error is an estimator of the test error, its variance may cause overfitting (Meanti et al., 2022). Hence, we will propose a similar method to (Meanti et al., 2022) that will avoid overfitting via utilizing the additional complexity regularization cost. Now, recall our original population level cost function for the second-stage (before simplification):

$$
\mathcal{L}^{2SR}(\varphi) = \mathbb{E}\big[\big(r(W, A) - \mathbb{E}[\varphi(Z, A)|W, A]\big)^2\big] + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}}} \qquad (44)
$$

Recall that our analysis has shown that φ must be in the form of

$$
\varphi = \sum_{i=1}^{m}\alpha_i\hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) + \alpha_{m+1}\frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{\substack{l=1 \\ l\neq j}}^{m}\hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j),
$$

if we optimize the corresponding sample loss in Equation (10). Now, suppose that we observe the noisy version of the target $r(W, A)$ (even though we do not observe the density-ratios, for theoretical analysis here we can see that Equation (44) is regression on the target variable $r(W, A)$). We write the sample-based counterpart of the loss in Equation (44) as follows:

$$
\hat{\mathcal{L}}_m^{\text{2SR},\epsilon}(\varphi) = \frac{1}{m}\|\boldsymbol{R}^\epsilon - \boldsymbol{L}\alpha\|^2 + \lambda_2\alpha^T\boldsymbol{N}\alpha \qquad (45)
$$

where $\boldsymbol{R}^\epsilon = \boldsymbol{R} + \boldsymbol{\epsilon}$ with $\text{Var}(\epsilon_i) = \sigma^2$ and $\mathbb{E}[\epsilon_i] = 0$ for all i, and

$$
\begin{aligned}
\boldsymbol{L} &= \begin{bmatrix}\boldsymbol{B}^T \boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]\frac{\mathbf{1}}{m}\end{bmatrix} \\
\boldsymbol{N} &= \begin{bmatrix}\boldsymbol{B}^T \boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T \boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]\frac{\mathbf{1}}{m} \\ (\frac{\mathbf{1}}{m})^T[\boldsymbol{B}^T \boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]^T & (\frac{\mathbf{1}}{m})^T\Big[\bar{\boldsymbol{B}}^T \boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big]\frac{\mathbf{1}}{m}\end{bmatrix}.
\end{aligned}
$$

Also, consider the noiseless case

$$
\hat{\mathcal{L}}_m^{\text{2SR}}(\varphi) = \frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2 + \lambda_2\alpha^T\boldsymbol{N}\alpha
$$

The optimum of $\hat{\mathcal{L}}^{\mathrm{2SR},\epsilon}(\varphi)$ can be written as

$$\begin{aligned}\alpha = \arg\min_{\alpha} \hat{\mathcal{L}}_m^{\mathrm{2SR},\epsilon}(\varphi) &= \big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{R}^{\epsilon} \\ &= \big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T(\boldsymbol{R} + \boldsymbol{\epsilon}).\end{aligned}$$

Now, consider the following expectation

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R}^{\epsilon} - \boldsymbol{L}\alpha\|^2\Big] &= \mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big] + \frac{2}{m}\mathbb{E}\Big[\langle \boldsymbol{R} - \boldsymbol{L}\alpha, \boldsymbol{\epsilon}\rangle\Big] + \sigma^2 \\ &= \mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big] + \frac{2}{m}\mathbb{E}\Big[\langle \boldsymbol{R} - \boldsymbol{L}\alpha, \boldsymbol{\epsilon}\rangle\Big] + \sigma^2 \\ &= \mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big] - \frac{2}{m}\mathbb{E}\Big[\langle \boldsymbol{L}\big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T(\boldsymbol{R} + \boldsymbol{\epsilon}), \boldsymbol{\epsilon}\rangle\Big] + \sigma^2 \\ &= \mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big] - \frac{2}{m}\mathbb{E}\Big[\langle \boldsymbol{L}\big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle\Big] + \sigma^2 \\ &= \mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big] - \frac{2\sigma^2}{m}\mathrm{Tr}\Big(\big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{L}\Big) + \sigma^2.\end{aligned}$$

Hence,

$$\mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big] = \mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R}^{\epsilon} - \boldsymbol{L}\alpha\|^2\Big] + \frac{2\sigma^2}{m}\mathrm{Tr}\Big(\big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{L}\Big) - \sigma^2.$$

Note that the term $\mathbb{E}\Big[\frac{1}{m}\|\boldsymbol{R} - \boldsymbol{L}\alpha\|^2\Big]$ is the *expected risk* in the noise free setup. Furthermore, the term $\frac{2\sigma^2}{m}\mathrm{Tr}\Big(\big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{L}\Big)$ is referred as *degrees of freedom* (or complexity) of the estimator, and it is a positive scalar. It penalizes the complex estimators. Having derived this, now consider the validation loss $\hat{\mathcal{L}}^{\mathrm{Val}}(\varphi)$ that we have previously derived. As we pointed out earlier, optimizing the regularization parameter λ_2 with respect to this validation loss can still lead to overfitting due to its variance. Hence, we propose to optimize the regularization parameter λ_2 with respect to the following surrogate cost that is both an upper bound on the validation error and penalizes the overly complex models:

$$\hat{\mathcal{L}}^{\mathrm{Val}}(\varphi) \leq \hat{\mathcal{L}}^{\mathrm{Val}}_{\sigma^2}(\varphi) \quad \forall \sigma \geq 0,$$

where

$$\begin{aligned}\hat{\mathcal{L}}^{\mathrm{Val}}_{\sigma^2}(\varphi) = \frac{1}{n}\begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\boldsymbol{C} \odot \boldsymbol{K}_{\tilde{A}A} \\ (\frac{\boldsymbol{1}}{m})^T\big(\bar{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\boldsymbol{C} \odot \boldsymbol{K}_{\tilde{A}A}\big)\end{bmatrix}\begin{bmatrix}\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}} & \big(\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}\big)\frac{\boldsymbol{1}}{m}\end{bmatrix}\begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix} \\ - 2\begin{bmatrix}\alpha_{1:m} \\ \alpha_{m+1}\end{bmatrix}^{\top}\begin{bmatrix}\big(\boldsymbol{B}^{\top}\boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\frac{\boldsymbol{1}}{n} \\ (\frac{\boldsymbol{1}}{m})^{\top}\big(\bar{\boldsymbol{B}}^{\top}\boldsymbol{K}_{ZZ}\bar{\boldsymbol{C}} \odot \boldsymbol{K}_{\tilde{A}A}\big)\frac{\boldsymbol{1}}{n}\end{bmatrix} + \frac{2\sigma^2}{m}\mathrm{Tr}\Big(\big(\boldsymbol{L}^T\boldsymbol{L} + m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{L}\Big) \qquad (46)\end{aligned}$$

Hence, we can tune λ_2 over a grid $\Lambda_2 \subset \mathbb{R}$ that minimizes the surrogate loss in Equation (46). One drawback of this approach is that we need to either estimate σ^2 or treat it as another hyperparameter. In our experiments, we opted to treat σ^2 as a hyperparameter. For the synthetic low-dimensional data and the legalized abortion and crime dataset, we set $\sigma^2 = 1$. In our dSprite and grade retention experiments, we used $\sigma^2 = 3$.

13.2.3 Tuning λ_2 Regularization Parameter in ATT Algorithm

Recall that in the second-stage regression for ATT estimation, we minimize the following loss:

$$\begin{aligned}\hat{\mathcal{L}}_m^{\mathrm{2SR}}(\varphi) &= \frac{1}{m}\sum_{i=1}^{m}\langle \varphi, \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a')\rangle^2 \\ &\quad - 2\frac{1}{m}\sum_{j=1}^{m}\sum_{\substack{i=1 \\ i \neq j}}^{m}\langle \varphi, \theta_i\mu_{Z|W,A}(\tilde{w}_i, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a')\rangle + \lambda_2\|\varphi\|^2_{\mathcal{H}_{\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{A}}}.\end{aligned}$$

We can compute the validation loss using the first-stage data $\{z_i, w_i, a_i\}_{i=1}^n$ similar to ATE algorithm with the following expression

$$\begin{aligned}\mathcal{L}^{\mathrm{Val}}(\varphi) &= \frac{1}{n}\sum_{i=1}^n \langle \varphi, \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a')\rangle^2 \\ &\quad - 2\frac{1}{n}\sum_{j=1}^n \sum_{\substack{i=1 \\ i\neq j}}^n \langle \varphi, \theta_i^{(2)} \hat{\mu}_{Z|W,A}(w_i, a_j) \otimes \phi_{\mathcal{A}}(a_j) \otimes \phi_{\mathcal{A}}(a')\rangle \end{aligned} \tag{47}$$

where $\theta_i^{(2)} = [(\boldsymbol{K}_{AA} + n\zeta^{(2)}\boldsymbol{I})^{-1}\boldsymbol{K}_{Aa'}]_i$. Furthermore, recall that the expression for φ that we have is

$$\varphi = \sum_{i=1}^m \alpha_i \hat{\mu}_{Z|W,A}(\tilde{w}_i, \tilde{a}_i) \otimes \phi_{\mathcal{A}}(\tilde{a}_i) \otimes \phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m}\sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a').$$

where the set $\{\alpha_i\}_{i=1}^{m+1}$ are the optimizer of the loss function in Equation (28). Now, we need to compute this validation loss in terms of matrix-vector multiplications. First, consider the following inner product:

$$\begin{aligned} &\Big\langle \varphi, \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a')\Big\rangle \\ &= \Bigg\langle \sum_{l=1}^m \alpha_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{x}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a') + \frac{\alpha_{m+1}}{m}\sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \\ &\hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a')\Bigg\rangle \\ &= \sum_{l=1}^m \alpha_l \Big\langle \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_l) \otimes \phi_{\mathcal{A}}(\tilde{a}_l) \otimes \phi_{\mathcal{A}}(a'), \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a')\Big\rangle \\ &+ \frac{\alpha_{m+1}}{m}\sum_{j=1}^m \sum_{\substack{l=1 \\ l\neq j}}^m \Big\langle \theta_l \hat{\mu}_{Z|W,A}(\tilde{w}_l, \tilde{a}_j) \otimes \phi_{\mathcal{A}}(\tilde{a}_j) \otimes \phi_{\mathcal{A}}(a'), \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a')\Big\rangle \\ &= \sum_{l=1}^m \alpha_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_l)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(w_i, a_i) k_{\mathcal{A}}(\tilde{a}_l, a_i) k_{\mathcal{A}}(a', a') \\ &+ \frac{\alpha_{m+1}}{m}\sum_{j=1}^m \Big(\sum_{\substack{l=1 \\ l\neq j}}^m \theta_l \boldsymbol{\beta}(\tilde{w}_l, \tilde{a}_j)\Big)^T \boldsymbol{K}_{ZZ} \boldsymbol{\beta}(w_i, a_i) k_{\mathcal{A}}(\tilde{a}_j, a_i) k_{\mathcal{A}}(a', a') \\ &= \Big[\big[\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}}\big]\alpha_{1:m}\Big]_i k_{\mathcal{A}}(a', a') + \alpha_{m+1}\Big[\big[\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}\big](\mathbf{1}/m)\Big]_i k_{\mathcal{A}}(a', a') \\ &= \left[\begin{bmatrix} [\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}} & [\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}](\mathbf{1}/m)] \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \right]_i k_{\mathcal{A}}(a', a') \end{aligned}$$

where

$$\boldsymbol{C} = \Big(\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda_1 \boldsymbol{I}\Big)^{-1}(\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA}).$$

As a result, the first component in Equation (47) is given by

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \Big\langle \varphi, \hat{\mu}_{Z|W,A}(w_i, a_i) \otimes \phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{A}}(a')\Big\rangle^2 \\ &= \frac{k_{\mathcal{A}}(a', a')^2}{n} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix}^T \begin{bmatrix} \big[\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}}\big]^T \\ (\frac{\mathbf{1}}{m})^T \big[\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}\big]^T \end{bmatrix} \begin{bmatrix} \boldsymbol{C}^T \boldsymbol{K}_{ZZ} \boldsymbol{B} \odot \boldsymbol{K}_{A\tilde{A}} & \big[\boldsymbol{C}^T \boldsymbol{K}_{ZZ} \tilde{\boldsymbol{B}} \odot \boldsymbol{K}_{A\tilde{A}}\big]\frac{\mathbf{1}}{m} \end{bmatrix} \begin{bmatrix} \alpha_{1:m} \\ \alpha_{m+1} \end{bmatrix} \end{aligned} \tag{48}$$

Next, for the second component in Equation (47), we note that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\Big\langle\varphi,\theta_j^{(2)}\hat{\mu}_{Z|W,A}(w_j,a_i)\otimes\phi_{\mathcal{A}}(a_i)\otimes\phi_{\mathcal{A}}(a')\Big\rangle\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{l=1}^{m}\alpha_l\Big\langle\hat{\mu}_{Z|W,A}(\tilde{w}_l,\tilde{a}_l)\otimes\phi_{\mathcal{A}}(\tilde{a}_l)\otimes\phi_{\mathcal{A}}(a'),\theta_j^{(2)}\hat{\mu}_{Z|W,A}(w_j,a_i)\otimes\phi_{\mathcal{A}}(a_i)\otimes\phi_{\mathcal{A}}(a')\Big\rangle\\
&+\frac{\alpha_{m+1}}{mn}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{r=1}^{m}\sum_{\substack{s=1\\ s\neq r}}^{m}\Big\langle\theta_s\hat{\mu}_{Z|W,A}(\tilde{w}_s,\tilde{a}_r)\otimes\phi_{\mathcal{A}}(\tilde{a}_r)\otimes\phi_{\mathcal{A}}(a'),\theta_j^{(2)}\hat{\mu}_{Z|W,A}(w_j,a_i)\otimes\phi_{\mathcal{A}}(a_i)\otimes\phi_{\mathcal{A}}(a')\Big\rangle\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{l=1}^{m}\alpha_l\boldsymbol{\beta}(\tilde{w}_l,\tilde{a}_l)^T\boldsymbol{K}_{ZZ}\theta_j^{(2)}\boldsymbol{\beta}(w_j,a_i)k_{\mathcal{A}}(\tilde{a}_l,a_i)k_{\mathcal{A}}(a',a')\\
&+\frac{\alpha_{m+1}}{mn}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{r=1}^{m}\sum_{\substack{s=1\\ s\neq r}}^{m}\theta_s\boldsymbol{\beta}(\tilde{w}_s,\tilde{a}_r)^T\boldsymbol{K}_{ZZ}\theta_j^{(2)}\boldsymbol{\beta}(w_j,a_i)k_{\mathcal{A}}(\tilde{a}_r,a_i)k_{\mathcal{A}}(a',a')\\
&=\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{m}\alpha_l\boldsymbol{\beta}(\tilde{w}_l,\tilde{a}_l)^T\boldsymbol{K}_{ZZ}\Big(\sum_{\substack{j=1\\ j\neq i}}^{n}\theta_j^{(2)}\boldsymbol{\beta}(w_j,a_i)\Big)k_{\mathcal{A}}(\tilde{a}_l,a_i)k_{\mathcal{A}}(a',a')\\
&+\frac{\alpha_{m+1}}{mn}\sum_{i=1}^{n}\sum_{r=1}^{m}\Big(\sum_{\substack{s=1\\ s\neq r}}^{m}\theta_s\boldsymbol{\beta}(\tilde{w}_s,\tilde{a}_r)\Big)^T\boldsymbol{K}_{ZZ}\Big(\sum_{\substack{j=1\\ j\neq i}}^{n}\theta_j^{(2)}\boldsymbol{\beta}(w_j,a_i)\Big)k_{\mathcal{A}}(\tilde{a}_r,a_i)k_{\mathcal{A}}(a',a')\\
&=\frac{1}{n}\alpha_{1:m}^T\Big[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}\Big]\mathbf{1}k_{\mathcal{A}}(a',a')+\alpha_{m+1}\frac{1}{mn}\mathbf{1}^T\Big[\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}\Big]\mathbf{1}k_{\mathcal{A}}(a',a')\\
&=\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}^T\begin{bmatrix}[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}]\frac{\mathbf{1}}{n}\\ (\frac{\mathbf{1}}{m})^T\Big[\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}\tilde{A}}\Big]\frac{\mathbf{1}}{n}\end{bmatrix}k_{\mathcal{A}}(a',a')
\end{aligned}
\tag{49}
$$

where

$$
\tilde{\boldsymbol{C}}_{:,j}=(\boldsymbol{K}_{WW}\odot\boldsymbol{K}_{AA}+n\lambda_1\boldsymbol{I})^{-1}\Big(\sum_{\substack{l=1\\ l\neq j}}^{n}\theta_l^{(2)}\boldsymbol{K}_{Ww_l}\odot\boldsymbol{K}_{Aa_j}\Big)
$$

Now, we are ready to combine our findings and write the loss function in terms of matrix-vector multiplications. Using Equations (48) and (49), the loss function can be expressed as

$$
\begin{aligned}
&\hat{\mathcal{L}}^{\text{Val}}(\varphi)=\frac{1}{n}\sum_{i=1}^{n}\langle\varphi,\hat{\mu}_{Z|W,A}(w_i,a_i)\otimes\phi_{\mathcal{A}}(a_i)\otimes\phi_{\mathcal{A}}(a')\rangle^2\\
&-2\frac{1}{n}\sum_{j=1}^{n}\sum_{\substack{i=1\\ i\neq j}}^{n}\langle\varphi,\theta_i^{(2)}\hat{\mu}_{Z|W,A}(w_i,a_j)\otimes\phi_{\mathcal{A}}(a_j)\otimes\phi_{\mathcal{A}}(a')\rangle\\
&=\frac{k_{\mathcal{A}}(a',a')^2}{n}\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\boldsymbol{B}\odot\boldsymbol{K}_{A\tilde{A}}\\ (\frac{\mathbf{1}}{m})^T\big[\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}}\odot\boldsymbol{K}_{A\tilde{A}}\big]^T\end{bmatrix}\begin{bmatrix}\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\boldsymbol{B}\odot\boldsymbol{K}_{A\tilde{A}} & [\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}}\odot\boldsymbol{K}_{A\tilde{A}}]\frac{\mathbf{1}}{m}\end{bmatrix}\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}\\
&-2\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}^T\begin{bmatrix}[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}]\frac{\mathbf{1}}{n}\\ (\frac{\mathbf{1}}{m})^T\Big[\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}\Big]\frac{\mathbf{1}}{n}\end{bmatrix}k_{\mathcal{A}}(a',a')
\end{aligned}
\tag{50}
$$

Similar to the tuning procedure of λ_2 in dose-response curve estimation, we can augment this validation loss with the complexity loss that will upper bound the hold-out loss while penalizing the overly complex models.

As a result, we can write the final surrogate loss to tune λ_2 as:

$$\begin{aligned}&\hat{\mathcal{L}}^{\mathrm{Val}}(\varphi)\\ &\leq \frac{k_{\mathcal{A}}(a',a')^2}{n}\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}^T\begin{bmatrix}\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\boldsymbol{B}\odot\boldsymbol{K}_{A\tilde{A}}\\ (\tfrac{\boldsymbol{1}}{m})^T\big[\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}}\odot\boldsymbol{K}_{A\tilde{A}}\big]^T\end{bmatrix}\begin{bmatrix}\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\boldsymbol{B}\odot\boldsymbol{K}_{A\tilde{A}} & \big[\boldsymbol{C}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}}\odot\boldsymbol{K}_{A\tilde{A}}\big]\frac{\boldsymbol{1}}{m}\end{bmatrix}\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}\\ &-2\begin{bmatrix}\alpha_{1:m}\\ \alpha_{m+1}\end{bmatrix}^T\begin{bmatrix}\big[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}\big]\frac{\boldsymbol{1}}{n}\\ (\tfrac{\boldsymbol{1}}{m})^T\Big[\tilde{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{C}}\odot\boldsymbol{K}_{\tilde{A}A}\Big]\frac{\boldsymbol{1}}{n}\end{bmatrix}k_{\mathcal{A}}(a',a')+\frac{2\sigma^2}{m}\mathrm{Tr}\Big(\big(\boldsymbol{L}^T\boldsymbol{L}+m\lambda_2\boldsymbol{I}\big)^{-1}\boldsymbol{L}^T\boldsymbol{L}\Big)\end{aligned}$$

where the matrix $\boldsymbol{L}$ is defined as

$$\boldsymbol{L}=\begin{bmatrix}\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\boldsymbol{B}\odot\boldsymbol{K}_{\tilde{A}\tilde{A}} & \big[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\tilde{\boldsymbol{B}}\odot\boldsymbol{K}_{\tilde{A}\tilde{A}}\big](\frac{\boldsymbol{1}}{m})\end{bmatrix}$$

In our ATT experiment that is illustrated in Figure (3), we used $\sigma^2 = 1$.

13.2.4 Tuning ζ Regularization Parameter in ATT Algorithm

In Algorithm (4.2), we estimate the conditional mean embedding $\mathbb{E}[\phi_{\mathcal{W}}(W)|A=a']$ by

$$\hat{\mathbb{E}}[\phi_{\mathcal{W}}(W)|A=a'] = \Phi_{\mathcal{W}}(\boldsymbol{K}_{\tilde{A}\tilde{A}}+m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{A}a'} = \sum_{i=1}^{m}\theta_i\phi_{\mathcal{W}}(\tilde{w}_i) = \Phi_{\mathcal{W}}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = (\boldsymbol{K}_{\tilde{A}\tilde{A}}+m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{A}a'}$. This is kernel ridge regression solution for measurements $\{\phi_{\mathcal{A}}(\tilde{a}_i)\}_{i=1}^m$ and targets $\{\phi_{\mathcal{W}}(\tilde{w}_i)\}_{i=1}^m$. Hence, one can use the LOOCV procedure presented in S.M. (13.2.1) to tune the regularization parameter in this estimator. Here, we will present another method that can be used for conditional mean embeddings that is provided in (Singh, 2023). Its proof can be found in the derivation of Algorithm 7 of (Singh, 2023).

Theorem 13.2 (Algorithm 7 in (Singh, 2023))**.** *Consider the conditional mean embedding*

$$\mathbb{E}[\phi_{\mathcal{W}}(W)|A=a'] = \mu_{W|A}(a') = \int\phi_{\mathcal{W}}(w)p(w|a')dw.$$

Sample based estimation using data $\{\tilde{w}_i,\tilde{a}_i\}_{i=1}^m$ for this conditional mean embedding is given by

$$\hat{\mu}_{W|A}(a') = \Phi_{\mathcal{W}}(\boldsymbol{K}_{\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}}+m\zeta\boldsymbol{I})^{-1}\boldsymbol{K}_{\tilde{\boldsymbol{A}}a'}$$

where $\Phi_{\mathcal{W}} = \begin{bmatrix}\phi_{\mathcal{W}}(\tilde{w}_1) & \phi_{\mathcal{W}}(\tilde{w}_2) & \dots & \phi_{\mathcal{W}}(\tilde{w}_m)\end{bmatrix}$, $[\boldsymbol{K}_{\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}}]_{ij} = k_{\mathcal{A}}(\tilde{a}_i,\tilde{a}_j)$, $[\boldsymbol{K}_{\tilde{\boldsymbol{A}}a'}]_i = k_{\mathcal{A}}(\tilde{a}_i,a')$ and $\boldsymbol{I}\in\mathbb{R}^{m\times m}$ is the identity matrix. The LOOCV loss for the conditional mean embedding estimation is given by

$$LOOCV_{\mu_{W|A}}(\zeta) = \frac{1}{m}Tr\Big(\boldsymbol{S}\big(\boldsymbol{K}_{\tilde{W}\tilde{W}}-2\boldsymbol{K}_{\tilde{W}\tilde{W}}\boldsymbol{R}^T+\boldsymbol{R}\boldsymbol{K}_{\tilde{W}\tilde{W}}\boldsymbol{R}^T\big)\Big) \tag{51}$$

where

$$\begin{aligned}[\boldsymbol{K}_{\tilde{W}\tilde{W}}]_{ij} &= k_{\mathcal{W}}(\tilde{w}_i,\tilde{w}_j)\\ \boldsymbol{R} &= \boldsymbol{K}_{\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}}(\boldsymbol{K}_{\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}}+m\zeta\boldsymbol{I})^{-1}\in\mathbb{R}^{m\times m}\\ \boldsymbol{S}\in\mathbb{R}^{m\times m}\ \ s.t.\ \ [\boldsymbol{S}]_{ij} &= \boldsymbol{1}[i=j]\left(\frac{1}{1-[\boldsymbol{R}]_{ij}}\right)^2.\end{aligned}$$

In our numerical experiments, we utilized Theorem (13.2) to tune regularization parameters ζ in ATT algorithm. In particular, we picked the regularization parameter that minimizes the LOOCV loss given in Equation (51) over a grid $\zeta\in\mathrm{Z}$ that is generated with a *logspace* with maximum and minimum values of 1.0 and 10^{-7}, respectively. We used 150 grid points in our numerical experiments.

13.3 Discussion on the Time Complexity of the Proposed Methods

Similar to kernel ridge regression, the complexity of our methods is governed by the matrix inversion. For simplicity, we consider the dose-response curve estimation in Algorithm (4.1). In the first-stage regression, the following matrix must be inverted:

$$\boldsymbol{K}_{WW} \odot \boldsymbol{K}_{AA} + n\lambda \boldsymbol{I} \in \mathbb{R}^{n \times n}.$$

This inversion operation has complexity of $O(n^3)$, making the first-stage sample size the limiting factor. Furthermore, to tune the regularization parameter λ_1 with LOOCV procedure, as discussed in S.M. (Sec. 13.2.1), this inversion must be performed for each grid point of $\lambda_1 \in \Lambda_1$.

In the second-stage, the following matrix needs to be inverted to obtain the optimizer coefficients $\{\alpha_i\}_{i=1}^{m+1}$ of Equation (18):

$$\frac{1}{m}\boldsymbol{L}^T\boldsymbol{L} + \lambda_2 \boldsymbol{N} \in \mathbb{R}^{(m+1)\times(m+1)}$$

where

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{B}^T\boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} \\ (\frac{\boldsymbol{1}}{m})^T\big[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\big]^T \end{bmatrix}^T \in \mathbb{R}^{m\times(m+1)},$$

$$\boldsymbol{N} = \begin{bmatrix} \boldsymbol{B}^T\boldsymbol{K}_{ZZ}\boldsymbol{B} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}} & [\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]\frac{\boldsymbol{1}}{m} \\ (\frac{\boldsymbol{1}}{m})^T[\boldsymbol{B}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}]^T & (\frac{\boldsymbol{1}}{m})^T\Big[\bar{\boldsymbol{B}}^T\boldsymbol{K}_{ZZ}\bar{\boldsymbol{B}} \odot \boldsymbol{K}_{\tilde{A}\tilde{A}}\Big]\frac{\boldsymbol{1}}{m} \end{bmatrix} \in \mathbb{R}^{(m+1)\times(m+1)},$$

as given in Algorithm (4.1). This inversion has complexity of $O((m+1)^3)$. Additionally, tuning the regularization parameter λ_2, as discussed in S.M. (Sec. 13.2.2), requires performing this inversion for each grid point for $\lambda_2 \in \Lambda_2$.

Finally, we note that the third-stage regression is another kernel ridge regression, where the matrix to be inverted (if only first-stage data is used) is

$$\boldsymbol{K}_{AA} + n\lambda_3 \boldsymbol{I} \in \mathbb{R}^{n\times n}.$$

As previously mentioned, data from either first-stage, second-stage, or a combination of both can be used in the third-stage regression. If the combination of first and second-stage data is used, the matrix to be inverted will have dimensions $(n+m)\times(n+m)$. Denoting by t the number of samples used for third-stage regression, the required inversion have complexity of $O(t^3)$. Therefore, to tune the regularization parameter λ_3 with LOOCV, as outlined in S.M. (Sec. 13.2.1), this inversion must be performed for each grid point of $\lambda_3 \in \Lambda_3$.

Overall, assuming $t = m + n$, the time complexity of our proposed method scales as $O(t^3)$. A similar analysis applies to our algorithm for conditional dose-response curve estimation.

13.4 Further Notes on Numerical Experiments

Our implementation code is available on GitHub link[2], which includes the instructions for reproducing the experiments presented in this paper.

For the tuning of the regularization parameters $(\lambda_1, \lambda_2, \lambda_3)$ (and ζ for ATT), we followed the procedures described in S.M. (Sec. 13.2.1), (Sec. 13.2.2), (Sec. 13.2.3) and (Sec. 13.2.4). In all experiments, the variables are normalized by subtracting mean and dividing by the standard deviation unless otherwise stated. The data were split uniformly into equal-sized first-stage and second-stage sets for the experiments. In the third-stage regression of our proposed methods, we used the combination of first and second-stage data.

13.5 Additional Numerical Experiments and Ablation Studies

13.5.1 Comparison of Our Approach and Outcome Bridge Function-Based Methods

A key question arising from our experiments in Section (6) is whether our method outperforms outcome bridge-based methods under specific conditions. To investigate this, we conducted synthetic experiments where one

[2] `https://github.com/BariscanBozkurt/Density-Ratio-Based-Proxy-Causal-Learning-without-Density-Ratios`

proxy variable was highly informative of confounders while the other was noisier. We adapted the data generation process from (Tsai et al., 2024, Appendix D.5) to construct various scenarios. We considered six scenarios where U follows different Beta distributions, and proxies Z and W vary in informativeness. The treatment and outcome variables were generated as follows:

Setting 1: $U \sim \text{Beta}(5,4)$, $W = g(U) + \mathcal{U}[0,1]$ where the function $g(x) = 0.8\frac{\exp(x)}{1+\exp(x)} + 0.1$ is applied elementwise, $Z = (1-U) \times Z_1 + U \times Z_2 + \mathcal{U}[0,100]$ where $Z_1 = \mathcal{N}(-1,0.1)$ and $Z_2 = \mathcal{N}(1,0.1)$, $A = 0.1U + 0.1Z + \mathcal{U}[0,1]$, and $Y = (2U-1) + \cos(1.5A)$.

Setting 2: $U \sim \text{Beta}(5,4)$, $Z = g(U) + \mathcal{U}[0,1]$, $W = (1-U) \times W_1 + U \times W_2 + \mathcal{U}[0,100]$ where $W_1 = \mathcal{N}(-1,0.1)$ and $W_2 = \mathcal{N}(1,0.1)$, $A = 0.1U + 0.1Z + \mathcal{U}[0,1]$, and $Y = (2U-1) + \cos(1.5A)$.

Setting 3: $U \sim \text{Beta}(8,4)$, $W = U + \mathcal{U}[0,1]$, $Z = g((1-U) \times Z_1 + U \times Z_2) + \mathcal{U}[0,100]$ where $Z_1 = \mathcal{N}(-1,0.1)$ and $Z_2 = \mathcal{N}(1,0.1)$, $A = 0.1U + 0.1Z + \mathcal{U}[0,1]$, and $Y = (2U-1) + \cos(1.5A)$.

Setting 4: $U \sim \text{Beta}(8,4)$, $Z = U + \mathcal{U}[0,1]$, $W = g((1-U) \times W_1 + U \times W_2) + \mathcal{U}[0,100]$ where $W_1 = \mathcal{N}(-1,0.1)$ and $W_2 = \mathcal{N}(1,0.1)$, $A = 0.1U + 0.1Z + \mathcal{U}[0,1]$, and $Y = (2U-1) + \cos(1.5A)$.

Setting 5: $U \sim \text{Beta}(3,5)$, $W = -U^2 + \mathcal{U}[0,1]$, $Z = g((1-U) \times Z_1 + U \times Z_2) + \mathcal{U}[0,100]$ where $Z_1 = \mathcal{N}(-1,0.1)$ and $Z_2 = \mathcal{N}(1,0.1)$, $A = 0.25\sqrt{|U|} - 0.2Z + \mathcal{U}[0,1]$, and $Y = 3W - 0.1A - \cos(0.5A + 5U)$.

Setting 6: $U \sim \text{Beta}(3,5)$, $Z = -U^2 + \mathcal{U}[0,1]$, $W = g((1-U) \times W_1 + U \times W_2 + \mathcal{U}[0,100])$ where $W_1 = \mathcal{N}(-1,0.1)$ and $W_2 = \mathcal{N}(1,0.1)$, $A = 0.25\sqrt{|U|} - 0.2Z + \mathcal{U}[0,1]$, and $Y = 3W - 2A - \cos(10A + 5U)$.

For each setting, we generated 1000 samples and evaluated our method against outcome bridge-based approaches over five runs, approximating the ground truth dose-response via Monte Carlo. Table (1) reports mean squared error and standard deviation across five independent realizations of each setting. Additionally, we compare the PCL algorithms to the oracle method Kernel-ATE, which directly uses the confounding variable U, in Table (1). Our method outperformed in odd-numbered settings where W was more informative, while outcome bridge-based methods excelled in even-numbered settings where Z was more informative. When the link between Z and U is highly noisy (or incomplete as in the next section), Assumption (3.7)—which ensures the existence of our treatment bridge function—is likely violated. Conversely, outcome bridge-based methods rely on this assumption for causal function identifiability. Thus, we hypothesize that our method is more robust when existence is challenged rather than identifiability. Meanwhile, we use Assumption (3.3) for identifiability, while KPV and KNC use it for outcome bridge function existence. This assumption is likely violated when the link between W and U is highly noisy (or incomplete), as in Settings 2, 4, and 6, where KPV and KNC outperform our method. We conjecture that violating our method's identifiability condition impacts performance more than violating bridge function existence.

Experimental results in S.M. (Sec. 13.5.2) also validates this hypothesis with the Job Corps dataset, where high-dimensional proxies were synthetically generated. Results reinforce the complementary strengths of treatment and outcome bridge-based methods. Understanding these trade-offs warrants deeper analysis, which we leave for future work. Nonetheless, our findings highlight the importance of further exploring treatment bridge-based approaches.

Table 1: Mean squared error for ablation studies with synthetic settings in S.M. (Sec. 13.5.1). We report mean ± standard deviation from $n = 5$ independent realizations of each setting.

	Kernel Alternative Proxy	**KNC**	**KPV**	**Kernel-ATE**
Setting 1	$\mathbf{0.00553 \pm 0.00069}$	0.29752 ± 0.08545	0.04184 ± 0.02661	0.00026 ± 0.00019
Setting 2	0.00932 ± 0.00529	0.01086 ± 0.00373	$\mathbf{0.00541 \pm 0.00217}$	0.00018 ± 0.00021
Setting 3	$\mathbf{0.00347 \pm 0.00104}$	0.25505 ± 0.09238	0.05122 ± 0.04610	0.00014 ± 0.00009
Setting 4	0.01532 ± 0.00548	0.01333 ± 0.00437	$\mathbf{0.00899 \pm 0.00491}$	0.00004 ± 0.00003
Setting 5	$\mathbf{0.01129 \pm 0.00823}$	0.03312 ± 0.02424	0.01982 ± 0.00885	0.01105 ± 0.00701
Setting 6	0.18436 ± 0.04889	0.09568 ± 0.02510	$\mathbf{0.05394 \pm 0.01516}$	0.00330 ± 0.00262

13.5.2 Dose-Response and Conditional Dose-Response Estimations in Job Corps Dataset

In this section, we conduct a new set of semi-synthetic experiments to compare our proposed method with the outcome bridge function-based algorithm based on the US Job Corps dataset (Schochet et al., 2008; Flores et al., 2012), adapted for the Proxy Causal Learning (PCL) setting. The US Job Corps Program is an educational intervention targeting disadvantaged youth. In this context, the continuous treatment variable A represents the total hours spent in academic or vocational training, while the continuous outcome variable Y corresponds to the proportion of weeks employed during the second year of training. Consistent with the setup in (Singh et al., 2023), the covariates $U \in \mathbb{R}^{65}$ include factors such as gender, ethnicity, age, language proficiency, education level, marital status, household size, and others. We obtained the dataset from the publicly available code of (Singh et al., 2023) (see `https://github.com/liyuan9988/KernelCausalFunction/tree/master`). To adapt this dataset to the PCL framework, we synthetically generated two proxy variables, Z and W, from U using the following settings:

Setting 1: $W = U + \epsilon$, $Z = g\left(U^{(1:20)}/\max\left(U^{(1:20)}\right)\right) + \nu$ where the function $g(x) = 0.8\frac{\exp(x)}{1+\exp(x)} + 0.1$ is the elementwise truncated logistic link function, $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i = 1,\ldots,65$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i = 1,\ldots,20$, and $U^{(1:20)}$ indicates taking the first 20 components of the vector U. The division operation and the max function are executed elementwise.

Setting 2: $Z = U + \epsilon$, $W = g\left(U^{(1:20)}/\max\left(U^{(1:20)}\right)\right) + \nu$ where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 3: $W = U + \epsilon$, $Z = g\left(U^{(20:40)}/\max\left(U^{(20:40)}\right)\right) + \nu$ where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 4: $Z = U + \epsilon$, $W = g\left(U^{(20:40)}/\max\left(U^{(20:40)}\right)\right) + \nu$ where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 5: $W = U + \epsilon$, $Z = g\left(U^{(40:60)}/\max\left(U^{(40:60)}\right)\right) + \nu$ where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 6: $Z = U + \epsilon$, $W = g\left(U^{(40:60)}/\max\left(U^{(40:60)}\right)\right) + \nu$ where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 7: $W = U + \epsilon$, $Z^{(1:20)} = g\left(U^{(1:20)}/\max\left(U^{(1:20)}\right)\right) + \nu$, $Z^{(21:65)} \sim \mathcal{N}(0,I)$, where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 8: $Z = U + \epsilon$, $W^{(1:20)} = g\left(U^{(1:20)}/\max\left(U^{(1:20)}\right)\right) + \nu$, $W^{(21:65)} \sim \mathcal{N}(0,I)$, where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 9: $W = U + \epsilon$, $Z^{(46:65)} = g\left(U^{(46:65)}/\max\left(U^{(46:65)}\right)\right) + \nu$, $Z^{(1:45)} \sim \mathcal{N}(0,I)$, where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 10: $Z = U + \epsilon$, $W^{(46:65)} = g\left(U^{(46:65)}/\max\left(U^{(46:65)}\right)\right) + \nu$, $W_{1:45} \sim \mathcal{N}(0,I)$, where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

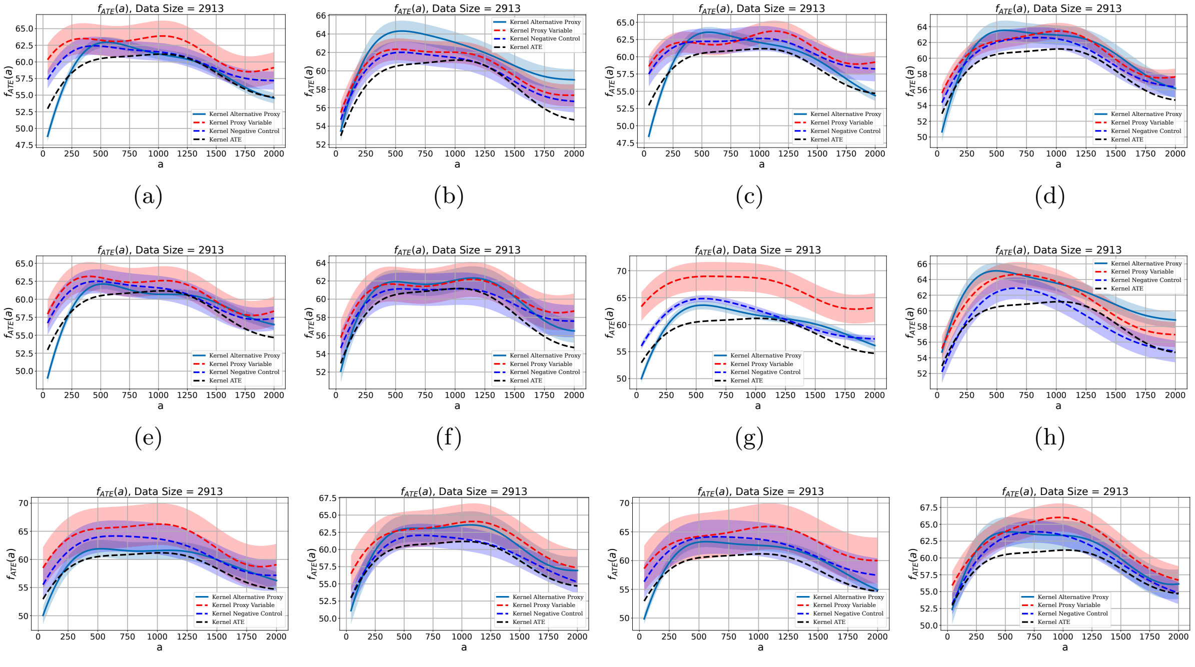
Setting 11: $W = U + \epsilon$, $Z^{(21:39)} = g\left(U^{(21:39)}/\max\left(U^{(21:39)}\right)\right) + \nu$, $Z^{(1:20)} \sim \mathcal{N}(0,I)$, $Z^{(40:65)} \sim \mathcal{N}(0,I)$, where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Setting 12: $Z = U + \epsilon$, $W^{(21:39)} = g\left(U^{(21:39)}/\max\left(U^{(21:39)}\right)\right) + \nu$, $W^{(1:20)} \sim \mathcal{N}(0,I)$, $W^{(40:65)} \sim \mathcal{N}(0,I)$, where $\epsilon^{(i)} \sim \mathcal{N}(0,1)\ \forall i$, $\nu^{(i)} \sim \mathcal{U}[-1,1]\ \forall i$.

Settings 1, 3, and 5 feature an incomplete link between the treatment proxy Z and the confounding variable U, combined with a nonlinearity function. These setups are likely to violate Assumption (3.7), which is crucial for the identifiability of KPV and KNC, as well as for the existence of the treatment bridge function in our method. Conversely, Settings 2, 4, and 6 feature an incomplete link between the outcome proxy W and the confounding variable U, along with a nonlinearity function. These setups are likely to violate Assumption (3.3), which is essential for the identifiability of our method and plays a role in establishing the existence of the outcome bridge function for KPV/KNC algorithms (Xu et al., 2021).

Figures (4a)-(4f) illustrate the simulation results, averaged over five runs with different two-stage splits, for Settings 1-6. We also compare our method against the Kernel-ATE algorithm (Singh et al., 2023), which uses U directly and serves as an oracle benchmark. For all algorithms, we employed Gaussian kernels with median length scale heuristics for all input variables. Each dimension of the input variables Z, W, and A (as well as U) is normalized by subtracting the mean and dividing by the standard deviation before being fed into our method, KPV, KNC, and Kernel-ATE. In settings where there is an incomplete link between Z and U, our method outperforms KPV and KNC, yielding results closer to the oracle method. In contrast, in settings where there is an incomplete link between W and U, KPV and/or KNC outperform our method. Consistent with our hypothesis in S.M. (Sec. 13.5.1), our method demonstrates better robustness when the existence of the bridge function is violated rather than when causal function identifiability is compromised, as seen in Settings 1, 3, and 5. Conversely, outcome bridge-based methods show better robustness when the existence of the outcome bridge function is violated rather than when causal function identifiability is compromised. These experimental results highlight the complementary strengths of treatment and outcome bridge-based methods under different assumptions.

We also investigate scenarios where the confounding variable U has a noisy link with one of the proxies, as detailed in Settings 7-12. Specifically, we generate a proxy variable of the same dimension as the confounding variable, but some of its dimensions consist entirely of noise. In Settings 7, 9, and 11, the proxy variable Z has a highly noisy link to the confounder U, in addition to nonlinearity, which is likely to violate Assumption (3.7). Similarly, in Settings 8, 10, and 12, the proxy variable W has a highly noisy link to the confounder U, which is likely to violate Assumption (3.3). Figures (4g)-(4l) present the estimation results for each setting, comparing treatment and outcome bridge-based methods against the oracle Kernel-ATE method. Our findings indicate that our method performs better in settings where Z has a noisy link with U, producing results closer to the oracle method. Conversely, outcome bridge-based algorithms perform better in settings where W has a noisy link with U, yielding estimates closer to the oracle method. In these noisy link experiments, we constructed the input kernel as a product of separate kernels for noisy and non-noisy dimensions. For instance, in Setting 7, we used the kernel $k_{\mathcal{Z}}(z_i, z_j) = k^{(1)}_{\mathcal{Z}}(z_i^{(1:20)}, z_j^{(1:20)}) \times k^{(2)}_{\mathcal{Z}}(z_i^{(21:65)}, z_j^{(21:65)})$ where $k^{(1)}_{\mathcal{Z}}$ and $k^{(2)}_{\mathcal{Z}}$ are Gaussian kernels with length scales determined using the median heuristic. Here, $z_i^{(1:20)}$ denotes the first 20 dimension of the i-th training variable z_i.



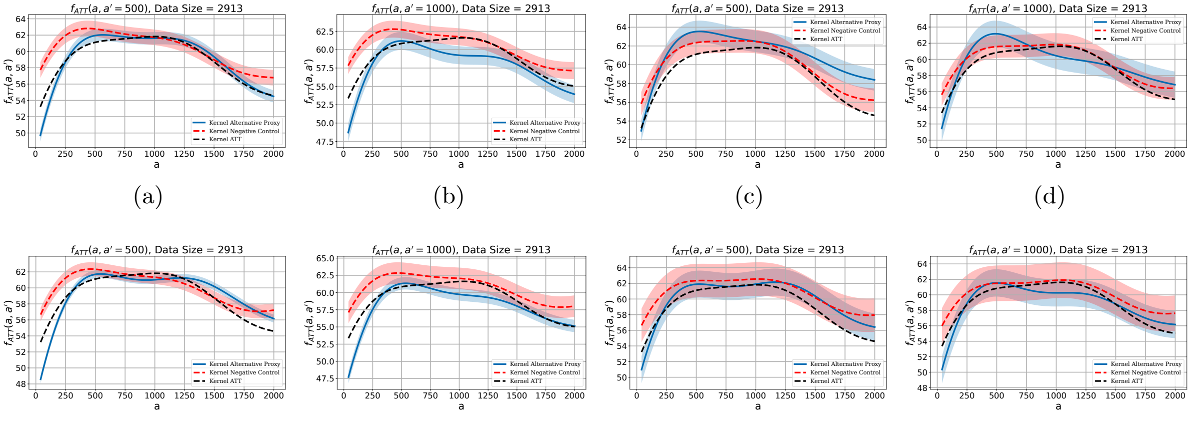
(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4: Dose-response estimation curves for the Job Corps experimental settings that are introduced in S.M. (Sec. 13.5.2). Panels (a)-(l) illustrate the estimation curves for our approach, KPV, KNC, and the oracle method Kernel-ATE across Settings 1-12, respectively.

To illustrate the conditional dose-response estimation capability of our proposed method in high dimensional settings, we also conduct experiments using the Job Corps dataset in Settings 1, 2, 5, and 6. Figures (5a) and (5b) show the ATT estimation results of our algorithm in comparison with the Kernel Negative Control method and the Kernel-ATT algorithm for $a' = 500$ and $a' = 1000$ in Setting 1, respectively. Figures (5c) and (5d) present the ATT estimation results for Setting 2. Figures (5e) and (5f) show the results for Setting 5, while Figures (5g) and (5h) provide results for Setting 6.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5: Conditional dose-response estimation curves for Job Corps experimental settings 1, 2, 5, and 6 that are introduced in S.M. (Sec. 13.5.2). Panels (a) and (b) show estimation curves for our approach, KNC, and the oracle method Kernel-ATT in Setting 1 for $a' = 500$ and $a' = 1000$, respectively. Panels (c) and (d) display the corresponding curves for Setting 2. Similarly, panels (e) and (f) illustrate the results for Setting 5, while panels (g) and (h) present those for Setting 6.

13.5.3 Ablation Study on the Effect of Bandwidth Selection of Gaussian Kernel

In this section, we present an ablation study to investigate the effect of kernel bandwidth selection on the performance of our dose-response curve estimation algorithm. Figure (6) illustrates the performance of our proposed method across different bandwidth selections for the kernels $k_{\mathcal{A}}(.,.)$, $k_{\mathcal{W}}(.,.)$, and $k_{\mathcal{Z}}(.,.)$. Specifically, Figure (6a) shows the performance our dose-response curve estimation algorithm for various bandwidth values for $k_{\mathcal{A}}(.,.)$ in the low-dimensional data generation setting (see Section (6)). The bandwidths of other kernels are set using median heuristic. Similarly, Figure (6b) and (6c) depicts the performance our method across different bandwidth selection for the kernels $k_{\mathcal{W}}(.,.)$ and $k_{\mathcal{Z}}(.,.)$, respectively.

We observe that while median heuristic does not always yield the best result, it generally produces robust or comparable results. Although one could perform a grid search on the kernel bandwidth to minimize the validation error in the second stage (see Equation (43)), this procedure introduces additional search complexity. Therefore, for simplicity, we opted to use median heuristic in our experiments.

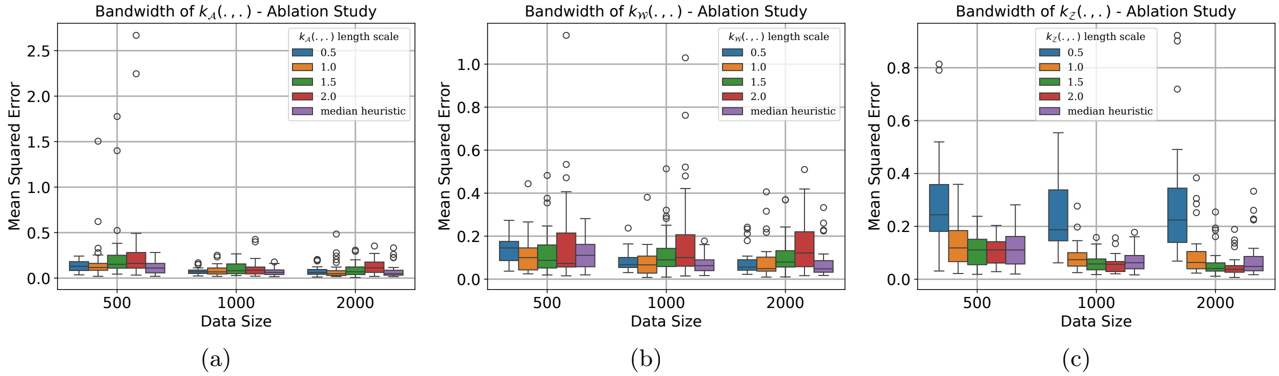
14 Identifiability of Dose-Response Curve in Discrete Case

Here, we additionally present the identification of dose-response curve in the discrete variable case. Assume that each space $\mathcal{F} \in \{\mathcal{Y}, \mathcal{W}, \mathcal{Z}, \mathcal{U}\}$ are discrete and $\mathcal{F} = \{1, 2, \dots, d_{\mathcal{F}}\}$. Note that the dose-response can be written in terms of matrix-vector products in this case:

$$\begin{aligned} f_{\text{ATE}}(a) &= \mathbb{E}_U[\mathbb{E}[Y|U, A = a]] = \sum_{i=1}^{d_{\mathcal{U}}} \mathbb{E}[Y|U = i, A = a]p(U = i) \\ &= \sum_{i=1}^{d_{\mathcal{U}}} \sum_{j=1}^{d_{\mathcal{Y}}} d_j p(Y = j|U = i)p(U = i) = \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y|U, A = \boldsymbol{a})\boldsymbol{P}(U) \end{aligned} \tag{52}$$

where $\boldsymbol{d}_{\mathcal{Y}} = \begin{bmatrix} 1 & 2 & \dots & d_{\mathcal{Y}} \end{bmatrix}^T \in \mathbb{R}^{d_{\mathcal{Y}}}$, and $\boldsymbol{P}(Y|U, A = a) \in \mathbb{R}^{d_{\mathcal{Y}} \times d_{\mathcal{U}}}$, $\boldsymbol{P}(U) \in \mathbb{R}^{d_{\mathcal{U}}}$ are probability matrices with

$$[\boldsymbol{P}(Y|U, A = \boldsymbol{a})]_{ij} = p(Y = i|U = j, A = \boldsymbol{a})$$



(a) (b) (c)

Figure 6: Ablation study on kernel bandwidth selection in the low-dimensional synthetic data experiment: (a) effect of the bandwidth of kernel $k_{\mathcal{A}}(.,.)$ on the performance, (b) effect of the bandwidth of kernel $k_{\mathcal{W}}(.,.)$ on the performance, (c) effect of the bandwidth of kernel $k_{\mathcal{Z}}(.,.)$ on the performance.

$$[\boldsymbol{P}(U)]_i = p(U = i).$$

Equation (52) cannot be computed directly since it involves the distribution of the unobserved confounder U. However, we can determine f_{ATE} by only the observable variables Y, W, Z. This approach is formalized in the following theorem.

Theorem 14.1. *Given the observed variables Y, W, Z from their corresponding discrete sets $\{\mathcal{Y}, \mathcal{W}, \mathcal{Z}\}$, the dose-response curve can be calculated by*

$$f_{ATE}(a) = \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y, Z|A = a)\boldsymbol{P}^{-T}(Z|W, A = a)\frac{1}{\boldsymbol{P}(A = a|W)^T}p(A = a) \tag{53}$$

where $\boldsymbol{P}(Y, Z|A = a) \in \mathbb{R}^{d_{\mathcal{Y}} \times d_{\mathcal{Z}}}$ and $\boldsymbol{P}(Z|W, A = a) \in \mathbb{R}^{d_{\mathcal{Z}} \times d_{\mathcal{W}}}$ are the probability matrices defined as

$$[\boldsymbol{P}(Y, Z|A = a)]_{ij} = p(Y = i, Z = j|A = a),$$
$$[\boldsymbol{P}(Z|W, A = a)]_{ij} = p(Z = i|W = j, A = a),$$

$\boldsymbol{P}^{-T}(Z|W, A = a)$ is the transpose of the inverse of the probability matrix $\boldsymbol{P}(Z|W, A = a)$, and

$$\frac{1}{\boldsymbol{P}(A = a|W)^T} = \begin{bmatrix} \frac{1}{p(A=a|W=1)} & \cdots & \frac{1}{p(A=a|W=d_{\mathcal{W}})} \end{bmatrix} \in \mathbb{R}^{1 \times d_{\mathcal{W}}}.$$

Note that every component in the Equation (53) is in terms of the observed variables and they can be estimated from the data.

Proof. We will prove the theorem in four steps:

Step 1: We will first prove that

$$\boldsymbol{P}(Z|W, A = a) = \boldsymbol{P}(Z|U, A = a)\boldsymbol{P}(U|W, A = a). \tag{54}$$

First consider the right-hand side of the Equation (54):

$$\begin{aligned}[\boldsymbol{P}(Z|U, A = a)\boldsymbol{P}(U|W, A = a)]_{ij} &= \sum_{k=1}^{d_{\mathcal{U}}} p(Z = i|U = k, A = a)p(U = k|W = j, A = a) \\ &= \sum_{k=1}^{d_{\mathcal{U}}} p(Z = i|U = k, A = a, W = j)p(U = k|W = j, A = a)\end{aligned}$$

(the above equality is due to Assumption (3.2))

$$\begin{aligned}
&= \sum_{k=1}^{d_{\mathcal{U}}} p(Z=i, U=k|A=a, W=j) \quad \text{(by Baye's Rule)} \\
&= p(Z=i|W=j, A=a) = [\boldsymbol{P}(Z|W, A=a)]_{ij}
\end{aligned}$$

and this verifies the Equation (54). It also implies that

$$\underbrace{\boldsymbol{P}(U|W, A=a)}_{\in \mathbb{R}^{d_{\mathcal{U}} \times d_{\mathcal{W}}}} = \underbrace{\boldsymbol{P}^{-1}(Z|U, A=a)}_{\in \mathbb{R}^{d_{\mathcal{U}} \times d_{\mathcal{Z}}}} \underbrace{\boldsymbol{P}(Z|W, A=a)}_{\in \mathbb{R}^{d_{\mathcal{Z}} \times d_{\mathcal{W}}}} \tag{55}$$

where $\boldsymbol{P}^{-1}(Z|U, A=a)$ is the (left) inverse of the probability matrix $\boldsymbol{P}(Z|U, A=a)$.

Step 2: Secondly, we want to show that

$$\boldsymbol{P}^{-T}(Z|U, A=a) \frac{p(A=a)}{\boldsymbol{P}(A=a|U)^T} = \boldsymbol{P}^{-T}(Z|W, A=a) \frac{p(A=a)}{\boldsymbol{P}(A=a|W)^T} \tag{56}$$

Equivalently, we will show that

$$\frac{1}{\boldsymbol{P}(A=a|U)} \boldsymbol{P}^{-1}(Z|U, A=a) = \frac{1}{\boldsymbol{P}(A=a|W)} \boldsymbol{P}^{-1}(Z|W, A=a)$$

which also implies that

$$\frac{1}{\boldsymbol{P}(A=a|U)} \boldsymbol{P}^{-1}(Z|U, A=a) \boldsymbol{P}(Z|W, A=a) = \frac{1}{\boldsymbol{P}(A=a|W)}. \tag{57}$$

Next, consider the left-hand side of the equation (57):

$$\begin{aligned}
&\Big[\frac{1}{\boldsymbol{P}(A=a|U)} \boldsymbol{P}^{-1}(Z|U, A=a) \boldsymbol{P}(Z|W, A=a)\Big]_i = \Big[\frac{1}{\boldsymbol{P}(A=a|U)} \boldsymbol{P}(U|W, A=a)\Big]_i \quad \text{(by Equation (55))} \\
&= \sum_{k=1}^{d_{\mathcal{U}}} \frac{1}{p(A=a|U=u)} p(U=k|W=i, A=a) \\
&= \sum_{k=1}^{d_{\mathcal{U}}} \frac{1}{p(A=a|U=u)} \frac{p(A=a|U=k, W=i) p(U=k|W=i)}{p(A=a|W=i)} \quad \text{(by Baye's Rule)} \\
&= \sum_{k=1}^{d_{\mathcal{U}}} \frac{1}{p(A=a|U=u)} \frac{p(A=a|U=k) p(U=k|W=i)}{p(A=a|W=i)} \quad \text{(since } W \perp A|U\text{, Assumption (3.2))} \\
&= \frac{1}{p(A=a|W=i)} \underbrace{\sum_{k=1}^{d_{\mathcal{U}}} p(U=k|W=i)}_{=1} = \frac{1}{p(A=a|W=i)} \\
&= \left[\frac{1}{\boldsymbol{P}(A=a|W)}\right]_i,
\end{aligned}$$

and that verifies the Equation (57). As a result, we proved that Equation (56) holds.

Step 3: We will further prove that

$$\boldsymbol{P}(Y|U, A=a) = \boldsymbol{P}(Y, Z|A=a) \boldsymbol{P}^{-T}(Z|U, A=a) \text{diag}\{\boldsymbol{P}(U|A=a)\}^{-1} \tag{58}$$

where

$$\text{diag}\{\boldsymbol{P}(U|A=a)\}^{-1} = \begin{bmatrix} \frac{1}{p(U=1|A=a)} & & & \Large 0 \\ & \frac{1}{p(U=2|A=a)} & & \\ & & \ddots & \\ \Large 0 & & & \frac{1}{p(U=d_{\mathcal{U}}|A=a)} \end{bmatrix} \in \mathbb{R}^{d_{\mathcal{U}} \times d_{\mathcal{U}}}$$

$$= p(A=a)\begin{bmatrix} \frac{1}{p(A=a|U=1)p(U=1)} & & & \mathbf{0} \\ & \frac{1}{p(A=a|U=2)p(U=2)} & & \\ & & \ddots & \\ \mathbf{0} & & & \frac{1}{p(A=a|U=d_{\mathcal{U}})p(U=d_{\mathcal{U}})} \end{bmatrix}$$

$$= p(A=a)\text{diag}\Big\{\frac{1}{\boldsymbol{P}(A=a|U)}\Big\}\text{diag}\Big\{\frac{1}{\boldsymbol{P}(U)}\Big\} \tag{59}$$

$$= p(A=a)\text{diag}\Big\{\boldsymbol{P}(A=a|U)\Big\}^{-1}\text{diag}\Big\{\boldsymbol{P}(U)\Big\}^{-1} \tag{60}$$

Next, consider

$$\Big[\text{diag}\{\boldsymbol{P}(U|A=a)\}\boldsymbol{P}^T(Z|U,A=a)\Big]_{ij} = \sum_{k=1}^{d_{\mathcal{U}}}\Big[\text{diag}\{\boldsymbol{P}(U|A=a)\}\Big]_{ik} p(Z=j|U=k,A=a)$$
$$= p(U=i|A=a)p(Z=j|U=i,A=a) \quad (\text{since } \Big[\text{diag}\{\boldsymbol{P}(U|A=a)\}\Big]_{ik} \text{ is nonzero iff } k=i)$$
$$= p(Z=j,U=i|A=a) \quad (\text{by Baye's Rule})$$

Hence, this illustrates that

$$\text{diag}\{\boldsymbol{P}(U|A=a)\}\boldsymbol{P}^T(Z|U,A=a) = \boldsymbol{P}^T(Z,U|A=a) \tag{61}$$

where $[\boldsymbol{P}(Z,U|A=a)]_{ij} = p(Z=i,U=j|A=a)$.

Furthermore, note that

$$\Big[\boldsymbol{P}(Y|U,A=a)\text{diag}\{\boldsymbol{P}(U|A=a)\}\boldsymbol{P}^T(Z|U,A=a)\Big]_{ij} = \Big[\boldsymbol{P}(Y|U,A=a)\boldsymbol{P}^T(Z,U|A=a)\Big]_{ij} \quad (\text{by Eq. (61)})$$
$$= \sum_{k=1}^{d_{\mathcal{U}}} p(Y=i|U=k,A=a)p(Z=j,U=k|A=a)$$
$$= \sum_{k=1}^{d_{\mathcal{U}}} p(Y=i|U=k,A=a)p(Z=j|U=k,A=a)p(U=k|A=a)$$
$$= \sum_{k=1}^{d_{\mathcal{U}}} p(Y=i,Z=j|U=k,A=a)p(U=k|A=a) \quad (\text{since } Y \perp Z|U,A, \text{ Assumption (3.2)})$$
$$= \sum_{k=1}^{d_{\mathcal{U}}} p(Y=i,Z=j,U=k|A=a) = p(Y=i,Z=j,U=k|A=a)$$
$$= \Big[\boldsymbol{P}(Y,Z|A=a)\Big]_{ij}.$$

Thus, we showed that

$$\boldsymbol{P}(Y|U,A=a)\text{diag}\{\boldsymbol{P}(U|A=a)\}\boldsymbol{P}^T(Z|U,A=a) = \boldsymbol{P}(Y,Z|A=a) \tag{62}$$

which also implies Equation (58). That is also equivalent to (due to Equation (60))

$$\boldsymbol{P}(Y|U,A=a) = \boldsymbol{P}(Y,Z|A=a)\boldsymbol{P}^{-T}(Z|U,A=a)p(A=a)\text{diag}\Big\{\frac{1}{\boldsymbol{P}(A=a|U)}\Big\}\text{diag}\Big\{\frac{1}{\boldsymbol{P}(U)}\Big\} \tag{63}$$

Step 4 (Combining all the steps above and finishing the proof):

Finally, consider the dose-response

$$f_{\text{ATE}}(a) = \mathbb{E}_U[\mathbb{E}[Y|U,A=a]] = \boldsymbol{d}_{\mathcal{Y}}^T\boldsymbol{P}(Y|U,A=a)\boldsymbol{P}(U)$$

$$= \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y, Z|A=a)\boldsymbol{P}^{-T}(Z|U, A=a)p(A=a)\text{diag}\left\{\frac{1}{\boldsymbol{P}(A=a|U)}\right\}\underbrace{\text{diag}\left\{\frac{1}{\boldsymbol{P}(U)}\right\}\boldsymbol{P}(U)}_{\boldsymbol{1}}$$

(the above equality is due to Equation (63))

$$= \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y, Z|A=a)\boldsymbol{P}^{-T}(Z|U, A=a)p(A=a)\text{diag}\left\{\frac{1}{\boldsymbol{P}(A=a|U)}\right\}\boldsymbol{1}$$

$$= \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y, Z|A=a)\underbrace{\boldsymbol{P}^{-T}(Z|U, A=a)p(A=a)\frac{1}{\boldsymbol{P}(A=a|U)^T}}_{\boldsymbol{P}^{-T}(Z|W,A=a)p(A=a)\frac{1}{\boldsymbol{P}(A=a|W)^T}}$$

$$= \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y, Z|A=a)\boldsymbol{P}^{-T}(Z|W, A=a)p(A=a)\frac{1}{\boldsymbol{P}(A=a|W)^T} \quad \text{(by Equation (56))}$$

As a result, the ATE functions can be calculated by

$$f_{\text{ATE}}(a) = \boldsymbol{d}_{\mathcal{Y}}^T \boldsymbol{P}(Y, Z|A=a)\boldsymbol{P}^{-T}(Z|W, A=a)p(A=a)\frac{1}{\boldsymbol{P}(A=a|W)^T}$$

where every component in the above equation is observed and the corresponding probability matrices can be estimated from the data. □

15 KERNEL ALTERNATIVE PROXY METHOD WITH OBSERVABLE CONFOUNDERS

In the section, we formulate the identifiability our proposed method when there exists observable confounding variables. In this setting, we consider the causal graph shown in Figure (7). In addition to the variables (A, Y, Z, W), we also assume that there exists observable confounding variables X. In this case, the structural functions of interest are defined as follows:

i-) Dose-response: $f_{\text{ATE}}(a) = \mathbb{E}[\mathbb{E}[Y|A=a, U, X]]$

ii-) Conditional dose-response: $f_{\text{ATT}}(a, a') = \mathbb{E}[\mathbb{E}[Y|A=a, U, X]|A=a']$

The conditional independence and completeness assumptions can be stated as follows:

Assumption 15.1. *We assume the following conditional independence statements: i-) $Y \perp Z|U, X, A$ (Conditional Independence for Y), ii-) $W \perp Z|U, X, A=a$ and $W \perp A|U, X$ (Conditional Independence for W).*

Assumption 15.2. *Let $\ell : \mathcal{U} \to \mathbb{R}$ be any square integrable function. We assume that the following conditions hold for all $a \in \mathcal{A}$, $x \in \mathcal{X}$:*

- $\mathbb{E}[\ell(U)|W=w, X=x, A=a] = 0 \quad \forall w \in \mathcal{W}$ *if and only if* $\ell(U) = 0 \quad p(U)-$*almost everywhere*
- $\mathbb{E}[\ell(U)|Z=z, X=x, A=a] = 0 \quad \forall z \in \mathcal{Z}$ *if and only if* $\ell(U) = 0 \quad p(U)-$*almost everywhere*

Next, we show the identifiability results of our proposed framework when there exist observable confounding variables $X \in \mathcal{X}$ for both dose-response and conditional dose-response case.

Theorem 15.3. *Assume there exists a* bridge function $\varphi_0(z, x, a)$ *that satisfies:*

$$\mathbb{E}[\varphi_0(Z, X, a)|W, X, A=a] = \frac{p(W|X)p(a)}{p(W, a|X)}.$$

Given the Assumptions (15.1) and (15.2), the dose-response curve is given by

$$f_{ATE}(a) = \mathbb{E}[Y\varphi_0(Z, X, a)|A=a]$$

Proof. Suppose that

$$\mathbb{E}[\varphi_0(Z, X, a)|W, X, A=a] = \frac{p(W|X)p(a)}{p(W, a|X)}.$$

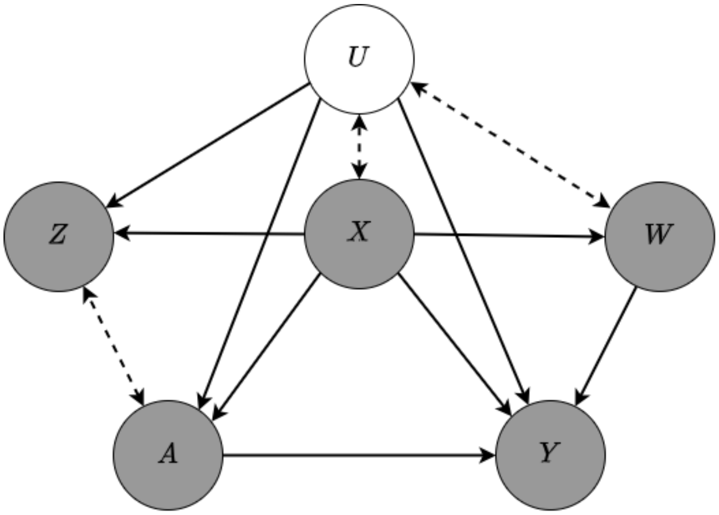


Figure 7: An instance of a Directed Acyclic Graph (DAG) for the PCL setting, which satisfies the required Assumption (15.1). In this graph, the gray circles denote the observed variables: A denotes the treatment, Y denotes the outcome, X denotes the additional observable confounding variables, Z denotes the treatment proxy, and W denotes the outcome proxy. The white circle denotes the unobserved confounding variable U. Bi-directional dotted arrows indicate that either direction in the DAG is possible, or that both variables may share a common ancestor.

Then, note the following,

$$\begin{aligned}\mathbb{E}[\varphi_0(Z,X,a)|W,X,A=a] &= \mathbb{E}_{U|W,X,A=a}[\mathbb{E}[\varphi_0(Z,X,a)|U,W,X,A=a]] \quad \text{(by Law of Total Expectations)}\\ &= \mathbb{E}_{U|W,X,A=a}[\mathbb{E}[\varphi_0(Z,X,a)|U,X,A=a]] \quad (\text{since } Z \perp W|U,X, \text{ Assumption (15.1)})\end{aligned} \tag{64}$$

Furthermore, note that

$$\begin{aligned}p(w|x) &= \int p(w|u,x)p(u|x)du\\ &= \int p(w|a,u,x)p(u|x)du \quad (\text{since } W \perp A|U,X, \text{ Assumption (15.1)})\\ &= \int \frac{p(u|w,a,x)p(w|a,x)}{p(u|a,x)}p(u|x)du \quad \text{(Baye's Rule)}\\ &= \int \frac{p(u|w,a,x)p(w,a|x)}{p(u,a|x)}p(u|x)du\\ &= p(w,a|x)\int \frac{p(u|x)}{p(u,a|x)}p(u|w,a,x)du\\ &= p(w,a|x)\mathbb{E}\left[\frac{p(U|x)}{p(U,a|x)}\middle|W=w,A=a,X=x\right]\end{aligned}$$

As a result,

$$\frac{p(w|X)}{p(w,a|X)} = \mathbb{E}\left[\frac{p(U|X)}{p(U,a|X)}\middle|W=w,A=a,X\right].$$

Hence,

$$\frac{p(w|X)p(a)}{p(w,a|X)} = \mathbb{E}\left[\frac{p(U|X)p(a)}{p(U,a|X)}\middle| W=w, A=a, X\right].$$

Using the assumption of the Theorem and Equation (64), we see that

$$\mathbb{E}_{U|W=w,X,A=a}[\mathbb{E}[\varphi_0(Z,X,a)|U,X,A=a]] = \mathbb{E}_{U|W=w,X,A=a}\left[\frac{p(U|X)p(a)}{p(U,a|X)}\right]$$

Therefore, due to the Assumption (15.2), we have

$$\mathbb{E}[\varphi_0(Z,X,a)|U,X,A=a] = \frac{p(U|X)p(a)}{p(U,a|X)} \quad \text{almost surely.} \tag{65}$$

Next, we observe that

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}[Y|A=a,U,X]] = \int \mathbb{E}[Y|A=a,u,x]p(u,x)dudx \\
&= \int \mathbb{E}[Y|A=a,u,x]\frac{p(u,x)p(a)}{p(u,a,x)}\frac{p(u,a,x)}{p(a)}dudx \\
&= \int \mathbb{E}[Y|A=a,u,x]\frac{p(u,x)p(a)}{p(u,a,x)}p(u,x|a)dudx \\
&= \mathbb{E}_{X,U|A=a}\left[\mathbb{E}[Y|A=a,U,X]\frac{p(U,X)p(a)}{p(U,a,X)}\right] \\
&= \mathbb{E}_{X,U|A=a}\left[\mathbb{E}[Y|A=a,U,X]\mathbb{E}[\varphi_0(Z,X,a)|U,X,A=a]\right] \quad \text{(by Equation 65)} \\
&= \mathbb{E}_{X,U|A=a}\left[\int yp(y|A=a,U,X)dy\int \varphi_0(z,X,a)p(z|U,X,A=a)dz\right] \\
&= \mathbb{E}_{X,U|A=a}\left[\int\int \varphi_0(z,X,a)y\underbrace{p(y|A=a,U,X,z)p(z|U,X,A=a)}_{p(y,z|A=a,U,X)}dydz\right] \text{(since } Y \perp Z|A=a,U,X\text{, Assump. (15.1))} \\
&= \int\int\int\int \varphi_0(z,x,a)y\underbrace{p(y,z|A=a,u,x)p(x,u|A=a)}_{p(u,x,y,z|A=a)}dydzdudx \\
&= \int\int\int \varphi_0(z,x,a)y\underbrace{\int p(u,x,y,z|A=a)du}_{p(y,z,x|A=a)}dydzdx \\
&= \int\int\int \varphi_0(z,x,a)yp(y,z,x|A=a)dydzdx = \mathbb{E}[Y\varphi_0(Z,X,a)|A=a].
\end{aligned}$$

As a result, we obtained

$$\mathbb{E}[Y\varphi_0(Z,X,a)|A=a] = \mathbb{E}_{X,U}[\mathbb{E}[Y|A=a,X,U]],$$

which indicates that $f_{\text{ATE}}(a) = \mathbb{E}[Y\varphi_0(Z,X,a)|A=a]$ and finishes the proof.

□

Theorem 15.4. *Assume there exists a* bridge function $\varphi_0(z,x,a,a')$ *that satisfies:*

$$\mathbb{E}[\varphi_0(Z,X,a,a')|W,X,A=a] = \frac{p(W,a'|X)p(a)}{p(W,a|X)p(a')}.$$

Given the Assumptions (15.1) and (15.2), the conditional dose-response curve is given by

$$f_{ATT}(a,a') = \mathbb{E}[Y\varphi_0(Z,X,a,a')|A=a]$$

Proof. First, observe the following

$$\mathbb{E}[\varphi_0(Z,X,a,a')|W,X,A=a] = \mathbb{E}_{U|W,X,A=a}\big[\mathbb{E}[\varphi_0(Z,X,a,a')|U,W,X,A=a]\big] \quad \text{(by Law of Total Expectations)}$$
$$= \mathbb{E}_{U|W,X,A=a}\big[\mathbb{E}[\varphi_0(Z,X,a,a')|U,X,A=a]\big] \quad \text{(since } Z \perp W|U,X,A=a\text{, Assumption (15.1)).} \tag{66}$$

Furthermore, note that

$$\begin{aligned} p(w,a'|x) &= \int p(w,a'|u,x)p(u|x)du = \int p(w|u,x)p(a'|u,x)p(u|x)du \quad \text{(since } W \perp A|U,X\text{, Assumption (15.1))} \\ &= \int p(w|a,u,x)p(a'|u,x)p(u|x)du \quad \text{(again due to } W \perp A|U,X\text{, Assumption (15.1))} \\ &= \int \frac{p(u|w,x,a)p(w|a,x)}{p(u|a,x)}p(a'|u,x)p(u|x)du \quad \text{(Baye's Rule)} \\ &= \int \frac{p(u|w,x,a)p(w,a|x)}{p(u,a|x)}p(u,a'|x)du = p(w,a|x)\int \frac{p(u,a'|x)}{p(u,a|x)}p(u|w,x,a)du. \end{aligned}$$

As a result,

$$\frac{p(w,a'|x)}{p(w,a|x)} = \int \frac{p(u,a'|x)}{p(u,a|x)}p(u|w,x,a)du.$$

Hence

$$\frac{p(w,a'|x)p(a)}{p(w,a|x)p(a')} = \mathbb{E}_{U|W=w,X=x,A=a}\left[\frac{p(U,a'|x)p(a)}{p(U,a|x)p(a')}\right]. \tag{67}$$

Recall that our assumption was

$$\mathbb{E}[\varphi_0(Z,X,a,a')|W,X,A=a] = \frac{p(W,a'|X)p(a)}{p(W,a|X)p(a)}$$

Thus, combining Equation (66) and Equation (67) yields

$$\mathbb{E}_{U|W=w,X,A=a}\Big[\mathbb{E}[\varphi_0(Z,X,a,a')|U,X,A=a]\Big] = \mathbb{E}_{U|W=w,X,A=a}\left[\frac{p(U,a'|X)p(a)}{p(U,a|X)p(a')}\right].$$

Using the completeness Assumption (15.2), we obtain

$$\mathbb{E}[\varphi_0(Z,X,a,a')|U,X,A=a] = \frac{p(U,a'|X)p(a)}{p(U,a|X)p(a')} \quad \text{almost surely} \tag{68}$$

Next, to obtain the ATT function, consider

$$\begin{aligned} &f_{\text{ATT}}(a,a') = \mathbb{E}_{U,X|A=a'}[\mathbb{E}[Y|A=a,U,X]] = \int \mathbb{E}[Y|A=a,u,x]p(u,x|A=a')du \\ &= \int \mathbb{E}[Y|A=a,u,x]\frac{p(u,x,A=a')}{p(a')}\frac{p(a)}{p(u,x,A=a)}\frac{p(u,x,A=a)}{p(a)}du \\ &= \int \mathbb{E}[Y|A=a,u,x]\frac{p(u,a'|x)p(a)}{p(u,a|x)p(a')}p(u,x|A=a)du \\ &= \mathbb{E}_{U,X|A=a}\Big[\mathbb{E}[Y|A=a,U,X]\mathbb{E}[\varphi_0(Z,X,a,a')|U,X,A=a]\Big] \quad \text{(by Equation (68))} \\ &= \mathbb{E}_{U,X|A=a}\left[\int yp(y|A=a,U,X)dy \int \varphi_0(z,X,a,a')p(z|A=a,U,X)dz\right] \end{aligned}$$

$$= \mathbb{E}_{U,X|A=a}\left[\int\int y\varphi_0(z,X,a,a')\underbrace{p(y|A=a,U,X,z)p(z|A=a,U,X)}_{p(y,z|A=a,U,X)}dydz\right]$$

$$= \mathbb{E}_{U,X|A=a}\left[\int\int y\varphi_0(z,X,a,a')p(y,z|A=a,U,X)dydz\right] \quad (Y \perp Z|U,X,A, \text{ Assumption (15.1) })$$

$$= \int\int\int\int y\varphi_0(z,x,a,a')\underbrace{p(y,z|A=a,U=u,X=x)p(u,x|A=a)}_{p(u,x,y,z|A=a)}dydzdudx$$

$$= \int\int\int y\varphi_0(z,x,a,a')\underbrace{\int p(u,x,y,z|A=a)du}_{p(y,z,x|A=a)}dxdydz$$

$$= \int\int y\varphi_0(z,x,a,a')p(y,z,x|A=a)dxdydz = \mathbb{E}[Y\varphi_0(Z,X,a,a')|A=a].$$

Hence, we have shown that $f_{\mathrm{ATT}}(a,a') = \mathbb{E}_{U,X|A=a'}[\mathbb{E}[Y|A=a,U,X]] = \mathbb{E}[Y\varphi_0(Z,X,a,a')|A=a]$. □