# A Completeness conditions

## A.1 Completeness condition for continuous and categorical confounder

The following two completeness conditions are necessary for the existence of solution for equation (1) and the consistency of causal effect inference should a solution exist. They are studied as equations (13) and (16) in Tchetgen Tchetgen et al. (2020).

1. For all $g \in \mathcal{L}^2_{P_U}$ and for any $a, x$, $\mathbb{E}[g(U)|a, x, z] = 0$ $P_Z - a.s.$ if and only if $g(U) = 0$ $P_U - a.s.$ This condition guarantees the viability of using the solution to (1) to consistently estimate the causal effect. Note that since $U$ is unobserved, this condition cannot be directly tested from observational data.

2. For all $g \in \mathcal{L}^2_{P_Z}$ and for any $a, x$, $\mathbb{E}[g(Z)|a, x, w] = 0$ $P_W - a.s.$ if and only if $g(Z) = 0$ $P_Z - a.s.$ This is a necessary condition for the existence of a solution to (1). With access to joint samples of $(a, x, w, z)$, in practice one can validate whether this condition holds and assess the quality of proxies $W, Z$ with respect to completeness condition. This assessment is beyond the scope of our study.

For a discrete confounder with categorical proxy variables, the combination of conditions 1 and 2 is equivalent to:

3. Both $W$ and $Z$ have at least as many categories as $U$.

4. For all $(a, x)$, the matrix $P$ where $P_{ij} = p(z_i|a, x, w_j)$ is invertible, with $z_i$ and $w_j$ denoting the $i$th and $j$th categories of $Z$ and $W$, respectively. Moreover, in the discrete case, this condition is necessary and sufficient for the solvability of Eq.1 as studied extensively in Miao et al. (2018), Tchetgen Tchetgen et al. (2020).

## A.2 Falsifying examples of the completeness condition

In this section we aim to provide intuition about the completeness conditions by giving examples of distributions which falsify them. For simplicity, we work with the completeness of $Z$ on $U$, which is the statement:

*$Z$ is complete for $U$ if and only if for all $g$ which is square-integrable, $\mathbb{E}[g(u)|z] = 0$ $\mathcal{P}_Z - a.s.$ if and only if $g(u) = 0$ $\mathcal{P}_U - a.s.$*

We proceed to provide examples in which the above statement fails to hold true.

- Trivial example. If $Z \perp\!\!\!\perp U$, then choose any non-zero square integrable $\tilde{g} \in \mathcal{L}^2(\mathcal{U})$ and define $g = \tilde{g} - \mathbb{E}[\tilde{g}(U)]$. Clearly $g \not\equiv 0$, but $\mathbb{E}[g(U)|Z] = \mathbb{E}[g(U)] = \mathbb{E}[\tilde{g}(U) - \mathbb{E}[\tilde{g}(U)]] = 0$

- Merely requiring that $Z$ and $U$ are dependent is not enough. Let $U = (X_1, X_2)$ and let $Z = (X_1, X_1)$ where $X_1 \perp\!\!\!\perp X_2$ and $X_1, X_2 \sim \mathcal{N}(0, 1)$. Thus $U$ and $Z$ are dependent. But let $g(U) = X_2$, then clearly $\mathbb{E}[g(U)|Z] = \mathbb{E}[X_2|Z] = \mathbb{E}[X_2] = 0$ for all $Z$ almost surely. Thus $Z$ is not complete for $U$.

- The reader might find the above two examples both trivial since they both require some component of $U$ to be independent of all components of $Z$. In the most general setting, the completeness condition is falsified if there is a $g \not\equiv 0 \in \mathcal{L}^2_{\mathcal{P}_Z}$ which is *orthogonal* to $\rho(u|z)$ for all values of $z$. This is equivalent to saying that:

$$\int_{\mathcal{U}} g(u)\rho(u|z)du = 0 \tag{17}$$

  or,

$$\int_{\mathcal{U}g^+} g^+(u)\rho(u|z)du = \int_{\mathcal{U}g^-} g^-(u)\rho(u|z)du \tag{18}$$

  $\mathcal{P}_Z - a.s.$, where $g^+$ and $g^-$ denotes the function or space restricted where $g$ is positive or negative, respectively. To see an example where this scenario can arise, and where all components of $Z$ are correlated with all components of $U$, consider the following. Let $U \sim \mathcal{N}(0, 1)$. $Z = f(U) + \mathcal{N}(0, 1) = |U| + \mathcal{N}(0, 1)$, where the added gaussian noise is independent of $U$. Let $g$ be a square integrable odd function, that is to say, $g(-x) = -g(x)$.

Then, we may examine the expectation of $g$ given $z$ as follows:

$$\begin{aligned}
\mathbb{E}[g(U)|z] &= \int_{-\infty}^{\infty} g(u)\rho(u|z)du && (19)\\
&= \int_{-\infty}^{0} g(u)\rho(u|z)du + \int_{0}^{\infty} g(u)\rho(u|z)du && (20)\\
&= \int_{\infty}^{0} -g(-v)\rho(-v|z)dv + \int_{0}^{\infty} g(u)\rho(u|z)du && (21)\\
&= \int_{0}^{\infty} g(-v)\rho(-v|z)dv + \int_{0}^{\infty} g(u)\rho(u|z)du && (22)\\
&= \int_{0}^{\infty} -g(v)\rho(-v|z)dv + \int_{0}^{\infty} g(u)\rho(u|z)du && (23)\\
&= \int_{0}^{\infty} -g(u)\rho(-u|z)du + \int_{0}^{\infty} g(u)\rho(u|z)du && (24)
\end{aligned}$$

where (21) is by taking substitution $v = -u$, (22) is swapping limit (23) is by oddness of $g$ and (24) is by renaming $v$ as $u$.

Now, $\rho(u|z)$ is symmetric in $U$, this can be seen by considering $\rho(u|z) = \frac{\rho(z|u)\rho(u)}{\rho(z)} \propto \rho(z|u)\rho(u)$.

$\rho(z|u)$ is symmetric in $u$ because $f(u) = |u|$ is symmetric; $\rho(u)$ is symmetric because it is a Gaussian; product of symmetric functions is symmetric.

Therefore,

$$(24) = \int_{0}^{\infty} -g(u)p(u|z)du + \int_{0}^{\infty} g(u)p(u|z)du = 0 \quad (25)$$

Thus no component of $Z$ is independent of $U$ but $Z$ is not complete for $U$.

Notice that in this case, we were able to construct such a $g$ because $f$ and $\rho$ have the same line of symmetry. Although this is an interesting example of falsification of the completeness condition, it is perhaps an unstable - i.e. we might be able to restore completeness if we slightly perturb the line of symmetry of $f$ and $\rho$.

**Remark 3.** *We note that although the completeness condition can be broken non-trivially by having a non-empty orthogonal set of $p(u|z)$ for almost all $z$, these cases might be unstable i.e. by slightly perturbing the joint distribution $\rho(u, z)$, so we hypothesize that the completeness condition is generically satisfied under mild conditions.*

# B Kernel Proxy Variable

## B.1 Notation

1. As $\mathcal{H}_\mathcal{P} \otimes \mathcal{H}_\mathcal{Q}$ is isometrically isomorphic to $\mathcal{H}_{\mathcal{PQ}}$, we use their features interchangeably, i.e. $\phi(p, q) = \phi(p) \otimes \phi(q)$.
2. $k(.,.)$ is a general notation for a kernel function, and $\phi(\cdot)$ denotes RKHS feature maps. To simplify notation, the argument of the kernel/feature map identifies it: for instance, $k(a, \cdot)$ and $\phi(a)$ denote the respective kernel and feature map on $\mathcal{A}$. We denote $K_{a\tilde{a}} := k(a, \tilde{a})$.
3. Kernel functions, their empirical estimates and their associated matrices are symmetric, i.e. $K_{ab} = K_{ba}$ and $K_{AA}^T = K_{AA}$. We use this property frequently in our proofs.

## B.2 Problem setting for RKHS-valued $h$

Recall that to estimate $h$ in (1), KPV aims at estimating $G_h(a, x, z)$ to minimize the empirical risk as:

$$\tilde{R}(h) = \mathbb{E}_{AXZY}\left[(Y - G_h(A, X, Z))^2\right], \text{ where } \quad G_h(a, x, z) := \int_{\mathcal{W}} h(a, x, w)\rho(w \,|\, a, x, z)dw$$

Since $h \in \mathcal{H}_{AXW}$ by Assumption 9, it follows from the reproducing property and the isometric isomorphism between Hilbert space of tensor products and product of Hilbert spaces that:

$$\begin{aligned} G_h(a,x,z) &:= \int_{\mathcal{W}} h(a,x,w)\rho(w \mid a,x,z)dw \\ &= \int_{\mathcal{W}} \langle h, \phi(a,x,w)\rangle_{\mathcal{H}_{AXW}} \rho(w \mid a,x,z)dw \\ &= \left\langle h, \int_{\mathcal{W}} \phi(a,x,w)\rho(w \mid a,x,z)dw \right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \\ &= \left\langle h, \int_{\mathcal{W}} [\phi(a)\otimes\phi(x)\otimes\phi(w)]\rho(w \mid a,x,z)dw \right\rangle_{\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{W}}} \\ &= \left\langle h, \phi(a)\otimes\phi(x)\otimes \int_{\mathcal{W}} \phi(w)\rho(w|a,x,z)dw \right\rangle_{\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{W}}} \\ &= \left\langle h, \phi(a)\otimes\phi(x)\otimes \mu_{W|a,x,z} \right\rangle_{\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{W}}} \end{aligned} \tag{26}$$

where $\mu_{W|a,x,z}$ denotes a conditional mean embedding of $\rho_{W\mid a,x,z}$, and we used the Bochner integrability (Steinwart & Christmann, 2008, Definition A.5.20) of the feature map $\phi(w)$ to take the expectation inside the dot product (this holds e.g. for bounded kernels). The regularised empirical risk minimization problem on $\{(\widetilde{a},\widetilde{x},\widetilde{z},\widetilde{y})_j\}_{j=1}^{m_2}$ can be expressed as:

$$\widetilde{\eta}_{AXW} = \underset{\eta\in\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}{\operatorname{argmin}} \ \widetilde{L}(\eta), \text{ where} \tag{27}$$

$$\widetilde{L}(\eta) = \frac{1}{m}\sum_{j=1}^{m_2}\left(\widetilde{y}_j - \left\langle \eta, \phi(\tilde{a}_j)\otimes\phi(\tilde{x}_j)\otimes\mu_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}\right\rangle_{\mathcal{H}_{\mathcal{A}}\otimes\mathcal{H}_{\mathcal{X}}\otimes\mathcal{H}_{\mathcal{W}}}\right)^2 + \lambda_2\|\eta\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$

with $\mu_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$ denoting the (true) conditional mean embedding of $\rho_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$. We will equivalently use the notation

$$\tilde{\eta}_{AXW}[\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\mu_{W|\widetilde{a},\widetilde{x},\widetilde{z}}] = \left\langle \tilde{\eta}_{AXW}, \phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\mu_{W|\widetilde{a},\widetilde{x},\widetilde{z}}\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$

to denote the evalation of $\tilde{\eta}_{AXW}$ at $\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\mu_{W|\widetilde{a},\widetilde{x},\widetilde{z}}$.

### B.3 A representer theorem expression for the empirical solution

**Lemma 3.** *Let $\widehat{\eta}_{AXW}$ be an empirical solution of (6), where the population conditional mean $\mu_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$ is replaced by an empirical estimate $\widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$ from (38). Then there exists $\alpha\in\mathbb{R}^{m_1\times m_2}$ such that:*

$$\widehat{\eta}_{AXW} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i). \tag{28}$$

*Proof.* Consider first the solution $\widetilde{\eta}_{AXW}$ of (27), where a population estimate of the conditional mean embedding $\mu_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$ is used in the first stage. By the representer theorem (Schölkopf et al., 2001), there exists $\gamma\in\mathbb{R}^{m_2}$ such that

$$\widetilde{\eta}_{AXW} = \sum_{j=1}^{m_2}\gamma_j\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\mu_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}. \tag{29}$$

In practice, we do not have access to the population embedding $\mu_{W|a,x,z}$. Thus, we substitute in an empirical estimate from (36),(38); see Stage 1 in Appendix B.4 for details. The empirical estimate of $\eta$ remains consistent under this replacement, and converges to its population estimate as *both* $m_1$ and $m_2$ increase (Theorem 2): see Appendix B.8 for the proof.

Substituting the empirical estimate $\widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$ from (38) in place of the population $\mu_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}$ in the empirical squared loss (27), then $\eta$ appears in a dot product with

$$\sum_{j=1}^{m_2}\phi(\tilde{a}_j)\otimes\phi(\tilde{x}_j)\otimes\underbrace{\left(\sum_{i=1}^{m_1}\Gamma_i(\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j)\phi(w_i)\right)}_{\widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Gamma_i(\tilde{a}_j,\widetilde{x}_j,\tilde{z}_j)\phi(\tilde{a}_j)\otimes\phi(\tilde{x}_j)\otimes\phi(w_i) \tag{30}$$

In other words, $\eta$ in the loss is evaluated at $m_1 \times m_2$ samples $(\tilde{a}_j, \tilde{x}_j, w_i)$. We know from the representer theorem (Schölkopf et al., 2001) that solutions $\widehat{\eta}_{AXW}$ are written in the span of $(\phi(\tilde{a}_j) \otimes \phi(\tilde{x}_j) \otimes \phi(w_i))$, $i \in \{1, \dots m_1\}$, $j \in \{1, \dots m_2\}$. The Gram matrix of these tensor sample features, appropriately rearranged, is an $(m_1 m_2) \times (m_1 m_2)$ matrix,

$$K_{\text{tot}} := K_{WW} \otimes (K_{AA} \odot K_{XX}),$$

where $K$ is the Kroenecker product. Assuming both $K_{WW}$ and $K_{AA} \odot K_{XX}$ have full rank, then by (Petersen & Pedersen, 2008, eq. 490), the rank of $K_{tot}$ is $m_1 m_2$ (in other words, the sample features used to express the representer theorem solution span a space of dimension $m_1 m_2$).

It is instructive to note that any empirical solution to (6) can hence be written as a linear combination of features of $\mathcal{H}_{\mathcal{AXW}}$, with features of $w$ from sample of the first stage $\{w_i\}_{i=1}^{m_1}$ and features of $a$ and $z$ from the second stage, $\{\tilde{a}_j, \tilde{x}_j\}_{j=1}^{m_2}$. □

**Remark:** We now provide further insight into the double sum form of (28). For simplicity, assume a single joint sample $\{(a, z, x, w, y)_i\}_{i=1}^n$, so that $m_1 = m_2 = n$. Given this sample, it might be tempting to write the Stage 2 KPV regression solution as single sum, rather than a double sum:

$$\widehat{\eta}_{\text{inc}} := \sum_{i=1}^{n} \alpha_i \phi(a_i) \otimes \phi(x_i) \otimes \phi(w_i). \tag{31}$$

Unfortunately, this solution is *incomplete*, and a double sum is needed for a correct solution. To see this, consider the subspace spanned by features making up the incomplete solution $(\phi(a_i) \otimes \phi(x_i) \otimes \phi(w_i))_{i=1}^n$ in (31). The Gram matrix for these sample features is

$$K_{\text{inc}} = K_{WW} \odot K_{AA} \odot K_{XX},$$

which has size $n \times n$, and rank at most $n$ (i.e., these features span a space of dimension at most $n$). Conseqently, the full Representer Theorem solution $\widehat{\eta}_{AXW}$ cannot be expressed in the form $\widehat{\eta}_{\text{inc}}$.

**Lemma 4.** *Let $\widehat{\eta}_{AXW}$ be expressed as (28). Then, its squared RKHS norm can be written as:*

$$\|\widehat{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{AXW}}} = \sum_{i,r=1}^{m_1} \sum_{j,t=1}^{m_2} \alpha_{ij}\, \alpha_{rt}\, K_{w_i w_r}\, K_{\widetilde{a}_j \widetilde{a}_t} K_{\widetilde{x}_j \widetilde{x}_t}. \tag{32}$$

*Proof.* By using the reproducing property and tensor product properties, we have:

$$\begin{aligned}
&\|\widehat{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{AXW}}} = \langle \widehat{\eta}_{AXW}, \widehat{\eta}_{AXW} \rangle_{\mathcal{F}_{\mathcal{AXW}}} \\
&= \left\langle \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \alpha_{ij} \phi(\widetilde{a}_j) \otimes \phi(\widetilde{x}_j) \otimes \phi(w_i), \sum_{r=1}^{m_1} \sum_{t=1}^{m_2} \alpha_{rt} \phi(\widetilde{a}_t) \otimes \phi(\widetilde{x}_t) \otimes \phi(w_r) \right\rangle_{\mathcal{F}_{\mathcal{AXW}}} \\
&= \sum_{i,r=1}^{m_1} \sum_{j,t=1}^{m_2} \alpha_{ij}\, \alpha_{rt}\, K_{w_i w_r}\, K_{\widetilde{a}_j \widetilde{a}_t} K_{\widetilde{x}_j \widetilde{x}_t},
\end{aligned} \tag{33}$$

where $\mathcal{F}_{\mathcal{AXW}}$ denotes the Frobenius (or Hilbert–Schmidt) inner product. In (33), we have used the known property of tensor product: $\langle a \otimes b, c \otimes d \rangle_{\mathcal{L}^2(\mathcal{H}_1, \mathcal{H}_2)} = \langle a, c \rangle_{\mathcal{H}_1} \otimes \langle b, d \rangle_{\mathcal{H}_2}$, where $\mathcal{L}^2(\mathcal{H}_1, \mathcal{H}_2)$ is the space of Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. Note that since $\widehat{\eta}_{AXW} = \phi(W)\alpha \otimes \phi(\widetilde{A}, \widetilde{X})$, its squared norm can also be written in trace form, as:

$$\begin{aligned}
&\|\widehat{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{AXW}}} = \langle \widehat{\eta}_{AXW}, \widehat{\eta}_{AXW} \rangle_{\mathcal{F}_{\mathcal{AXW}}} \\
&= Tr\left\{ \alpha^T K_{WW} \alpha (K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}) \right\}
\end{aligned} \tag{34}$$

using the connection between the *Trace* and *Hilbert Schmidt* or *Frobenius norm* and the reproducing property. □

### B.4 Kernel Proxy Variable Algorithm

In the previous section, we obtained a representer theorem for the form of the solution to (27), in the event that an empirical estimate $\widehat{\mu}_{W|a,x,z}$ is used for the mean embedding $\mu_{W|a,x,z}$, the (true) conditional mean embedding of $\rho_{W|a,x,z}$.

We have two goals for the present section: first, to provide an explicit form for $\widehat{\mu}_{W|a,x,z}$ (*Stage 1*). Second, in (*Stage 2*), to learn $\widehat{\eta}_{AXW}$, using the empirical embedding $\widehat{\mu}_{W|a,x,z}$ learned in stage 1. Theorem 2 show that the empirical estimate of $\eta$ remains consistent under this replacement and converges to its true value at population level, see Appendix B.8 for details. Consistent with the two-stages of the algorithm, we assume that the sample is divided into two sub-samples of size $m_1$ and $m_2$, i.e., $\{(a,x,z,w)_i\}_{i=1}^{m_1}$ and $\{(\widetilde{a},\widetilde{x},\widetilde{y},\widetilde{z})_j\}_{j=1}^{m_2}$.

**Stage 1.** Estimating Conditional mean embedding operator $\widehat{C}_{W|A,X,Z}$ from the first sample, $\{(a,x,z,w)_i\}_{i=1}^{m_1}$.

As stated in Assumption 7, $k(a,),k(x,),k(w,)$ and $k(z,)$ are characteristic kernels, and are continuous, bounded by $\kappa>0$, and $\mathbb{E}[\sqrt{k(\cdot,\cdot)}]<\infty$. We may define the *conditional mean embedding operator* as in Song et al. (2009):

$$C_{W|A,X,Z}:\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}\mapsto\mathcal{H}_{\mathcal{W}},\quad C_{W|A,X,Z}(\phi(a)\otimes\phi(x)\otimes\phi(z))=\mathbb{E}[\Psi(W)|a,x,z].$$

Following Singh et al. (2019, Theorem 1), it can be shown that

$$\widehat{C}_{W|A,X,Z}=\Psi(W)\left[K_{AA}\odot K_{XX}\odot K_{ZZ}+m_1\lambda_1 I_{m_1}\right]^{-1}\left[\Phi_{\mathcal{A}\mathcal{X}\mathcal{Z}}(A,X,Z)\right]^T, \tag{35}$$

where $K_{AA},K_{XX}$ and $K_{ZZ}$ are $m_1\times m_1$ kernel matrices and $\Psi(W)$ is a vector of $m_1$ columns, with $\phi(w_i)$ in its $i$th column. By definition (Song et al., 2009), $\widehat{\mu}_{W|a,x,z}:=\widehat{C}_{W|A,X,Z}\left(\phi(a)\otimes\phi(x)\otimes\phi(z)\right)$, and therefore

$$\begin{aligned}\widehat{\mu}_{W|a,x,z}&=\left[\Psi(W)\left[K_{AA}\odot K_{XX}\odot K_{ZZ}+m_1\lambda_1\right]^{-1}\left[\Phi(A)\otimes\Phi_{\mathcal{X}}(X)\otimes\Upsilon(Z)\right]^T\right]\left(\phi(a)\otimes\phi(x)\otimes\phi(z)\right)\\&=\Psi(W)\Gamma(a,x,z),\end{aligned} \tag{36}$$

where we applied the reproducing property and used isometric isomorphism between Hilbert space of tensor products and product of Hilbert spaces, i.e. $\Phi_{\mathcal{A}\otimes\mathcal{X}\otimes\mathcal{Z}}(A,X,Z)=\Phi(A)\otimes\Phi_{\mathcal{X}}(X)\otimes\Upsilon(Z)$. We defined $\Gamma(a,x,z)$ as a column matrix with $m_1$ rows :

$$\Gamma(a,x,z)=\left[\mathcal{K}_{AXZ}+m_1\lambda_1\right]^{-1}\mathcal{K}_{axz} \tag{37}$$

where $\mathcal{K}_{AXZ}=K_{AA}\odot K_{XX}\odot K_{ZZ}$ and $\mathcal{K}_{axz}=K_{Aa}\odot K_{Xx}\odot K_{Zz}$ are a $m_1\times m_1$ matrix and a column matrix with $m_1$ rows, respectively. Note that for any given $(a,x,z)$, $\mu_{W|a,x,z}\in$ *Span*$\{\Psi(W)\}$, and its empirical estimate can be expressed as

$$\widehat{\mu}_{W|a,x,z}=\sum_{i=1}^{m_1}\Gamma_i(a,x,z)\phi(w_i),\qquad\forall w_i\in\{(a,x,z,w)\}_{i=1}^{m_1}. \tag{38}$$

We now detail the second step where we use $\widehat{\mu}_{W|a,x,z}$ to learn the operator $\eta$ to minimize the empirical loss (6).

**Stage 2.** Expressing $\widehat{\eta}_{AXW}$ using $\{(\widetilde{a},\widetilde{x},\widetilde{y},\widetilde{z})_j\}_{j=1}^{m_2}$ and Stage 1.

It follows from (28) that for any $\{(\widetilde{a},\widetilde{x},\widetilde{z})_j\}_{j=1}^{m_2}\in(\mathcal{A},\mathcal{X},\mathcal{Z})$,

$$\begin{aligned}&\left\langle\widehat{\eta}_{AXW},\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\widehat{\mu}_{W|\widetilde{a},\widetilde{x},\widetilde{z}}\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\\&=\left\langle\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i),\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\widehat{\mu}_{W|\widetilde{a},\widetilde{x},\widetilde{z}}\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\\&=\left\langle\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i),\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\left\{\sum_{s=1}^{m_1}\Gamma_s(\widetilde{a},\widetilde{x},\widetilde{z})\phi(w_s)\right\}\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\\&=\left\langle\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i),\sum_{s=1}^{m_1}\Gamma_s(\widetilde{a},\widetilde{x},\widetilde{z})\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\phi(w_s)\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\\&=\sum_{i=1}^{m_1}\sum_{s=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\Gamma_s(\widetilde{a},\widetilde{x},\widetilde{z})\left\langle\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i),\phi(\widetilde{a})\otimes\phi(\widetilde{x})\otimes\phi(w_s)\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\\&=\sum_{i=1}^{m_1}\sum_{s=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\Gamma_s(\widetilde{a},\widetilde{x},\widetilde{z})k(w_i,w_s)k(\widetilde{a}_j,\widetilde{a})k(\widetilde{x}_j,\widetilde{x})\end{aligned} \tag{39}$$

where $k(w_i, w_s), k(\widetilde{a}_j, \widetilde{a})$ and $k(\widetilde{x}_j, \widetilde{x})$ denote associated kernels for variables $w, a$ and $x$. The second equation follows from (4). Substituting the expression of $\Gamma_s(\widetilde{a}, \widetilde{x}, \widetilde{z})$ from (37), we have for any $(a, x, z)$:

$$\begin{aligned}\widehat{\eta}_{AXW}[\phi(\widetilde{a}) \otimes \phi(\widetilde{x}) \otimes \widehat{\mu}_{W|\widetilde{a},\widetilde{x},\widetilde{z}}] &= \left\langle \widehat{\eta}_{AXW}, \phi(\widetilde{a}) \otimes \phi(\widetilde{x}) \otimes \widehat{\mu}_{W|\widetilde{a},\widetilde{x},\widetilde{z}} \right\rangle_{\mathcal{H}_{\mathcal{AXW}}} \\ &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{s=1}^{m_1} \alpha_{ij} K_{w_i w_s} \left\{ \left[K_{AA} \odot K_{XX} \odot K_{ZZ} + m_1\lambda_1\right]^{-1} \left[K_{A\widetilde{a}} \odot K_{X\widetilde{x}} \odot K_{Z\widetilde{z}}\right] \right\}_s \left[K_{\widetilde{a}_j\widetilde{a}} \odot K_{\widetilde{x}_j\widetilde{x}}\right] \end{aligned} \tag{40}$$

with $K_{AA}, K_{XX}$ and $K_{ZZ}$, $m_1 \times m_1$ matrices of empirical kernels of $A, X$ and $Z$ estimated from sample 1.

Equation (40) can be written in matrix format as:

$$\begin{aligned}&\left\langle \widehat{\eta}_{AXW}, \phi(\widetilde{a}) \otimes \phi(\widetilde{x}) \otimes \widehat{\mu}_{W|\widetilde{a},\widetilde{x},\widetilde{z}} \right\rangle_{\mathcal{H}_{\mathcal{AXW}}} = \\ &\qquad \left[K_{\widetilde{a}\widetilde{A}} \odot K_{\widetilde{x}\widetilde{X}}\right] \alpha^T K_{WW} \{ \left[K_{AA} \odot K_{XX} \odot K_{ZZ} + m_1\lambda_w\right]^{-1} \left[K_{A\widetilde{a}} \odot K_{X\widetilde{x}} \odot K_{Z\widetilde{z}}\right] \}\end{aligned}$$

This format will be convenient when deriving the closed-form solution for ERM (6).

Finally, combining from eq. (40) and (32), the ERM (6) can be written as a minimization over $\widehat{\alpha} \in \mathbb{R}^{m_1 \times m_2}$:

$$\widehat{\alpha} = \underset{\alpha \in \mathbb{R}^{m_1 \times m_2}}{\operatorname{argmin}} \widehat{L}(\alpha), \qquad \widehat{L}(\alpha) = \frac{1}{m_2}\sum_{q=1}^{m_2}\left(\widetilde{y}^q - \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij} A_{ij}^q\right)^2 + \lambda_2 \sum_{i,s=1}^{m_1}\sum_{j,t=1}^{m_2} \alpha_{ij}\,\alpha_{st}\, B_{ij}^{st}, \tag{41}$$

denoting $A_{ij}^q$ and $B_{ij}^{st}$ as

$$\begin{aligned} A_{ij}^q &= \left\{K_{w_i W}\left[K_{AA} \odot K_{XX} \odot K_{ZZ} + m_1\lambda_1\right]^{-1}\left[K_{A\widetilde{a}_q} \odot K_{X\widetilde{x}_q} \odot K_{Z\widetilde{z}_q}\right]\right\}\left[K_{\widetilde{a}_j\widetilde{a}_q} \odot K_{\widetilde{x}_j\widetilde{x}_q}\right], \\ B_{ij}^{st} &= K_{w_i w_s}\, K_{\widetilde{a}_j\widetilde{a}_t} K_{\widetilde{x}_j\widetilde{x}_t}. \end{aligned}$$

A solution $\widehat{\alpha} = [\widehat{\alpha}_{ij}]_{m_1 \times m_2}$ can be derived by solving $\frac{\partial \widehat{L}(\alpha)}{\partial \alpha} = 0$. As such, $\widehat{\alpha}$ is the solution to the system of the $m_1 \times m_2$ linear equations,

$$\forall (i,j) \in m_1 \times m_2 : \qquad \sum_q^{m_2} \widetilde{y}^q\, A_{ij}^q = \sum_s^{m_1}\sum_t^{m_2} \hat{\alpha}_{st}\left[\sum_q^{m_2} A_{ij}^q\, A_{st}^q \;+\; m_2\,\lambda_2\, B_{ij}^{st}\right] \tag{42}$$

**Remark 4.** *While the system of equations (42) is linear, deriving the solution requires inversion of a $m_1m_2 \times m_1m_2$ matrix. With a memory requirement of complexity $\mathcal{O}(m_1m_2)^2$ and $\mathcal{O}(m_1m_2)^3$, respectively, this is not possible in practice for even moderate sample sizes. We provide a computationally efficient solution in the next section.*

### B.5 Efficient closed-form solution for $\widehat{\eta}_{AXW}$: Proof of Proposition 2

As we explained in the previous section, deriving a solution for $\alpha$ – and consequently empirical estimate of $\eta$ – involves inverting a matrix $\in \mathbb{R}^{m_1m_2 \times m_1m_2}$, which is too computationally expensive for most applications. In this section, we propose an efficient method for finding $\widehat{\alpha}$. First, we vectorize the empirical loss (6); second, we employ a *Woodbudy Matrix Identity*.

#### B.5.1 VECTORIZING ERM (6)

The empirical risk, $\widehat{L}(\alpha)$, is a scalar, and it is a function of $\alpha$, a matrix. The idea of this section is to vectorise $\alpha$ as $v := vec(\alpha)$, and express empirical loss as a function of $v$. Naturally, this requires manipulation both the total expected loss $\mathbb{E}(Y - \widehat{Y})^2$ and the regularisation. In following sections, we show how to express these terms as functions of $v := vec(\alpha)$.

**Lemma 5.** *Vectorizing $\widehat{\alpha}$ as $\widehat{v} := vec(\widehat{\alpha})$, the ERM (41) can be expressed as:*

$$\widehat{v} = \underset{v \in \mathbb{R}^{m_1m_2}}{\operatorname{argmin}} \widehat{L}(v), \quad \widehat{L}(v) = \frac{1}{m_2}\|Y - v^T D\|_2^2 + \lambda_2 v^T E v \tag{43}$$

*Where:*

$$C = K_{WW}\Gamma(\widetilde{A},\widetilde{X},\widetilde{Z}) = K_{WW}\left[K_{AA}\odot K_{XX}\odot K_{ZZ}+m_1\lambda_w\right]^{-1}\left[K_{A\widetilde{A}}\odot K_{X\widetilde{X}}\odot K_{Z\widetilde{Z}}\right] \quad \in \mathbb{R}^{m_1\times m_2} \tag{44}$$

$$D = C\overline{\otimes}\left[K_{\widetilde{A}\widetilde{A}}\odot K_{\widetilde{X}\widetilde{X}}\right] \quad \in \mathbb{R}^{(m_1m_2)\times m_2} \tag{45}$$

$$E = K_{WW}\otimes\left(K_{\widetilde{A}\widetilde{A}}\odot K_{\widetilde{X}\widetilde{X}}\right) \quad \in \mathbb{R}^{m_1m_2\times m_1m_2}\quad , \tag{46}$$

*with $\otimes$ and $\overline{\otimes}$ representing tensor (Kronecker) product and tensor product of associated columns of matrices with the same number of columns, respectively. Vectorization is defined with regards to the rows of a matrix.*

*Proof.* The proof proceeds in two steps. Assume $\eta$ can be written as :

$$\eta = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i), \tag{47}$$

for $\alpha\in\mathbb{R}^{m_1\times m_2}$. We first show vectorized form of $\sum_{q=1}^{m}(y_q-\eta[\phi(a_q)\otimes\phi(x_q)\otimes\widehat{\mu}_{W|a_q,x_q,z_q}])^2$, and then that of the regularization term $\|\eta\|^2_{\mathcal{H}_{\mathcal{A}\times\mathcal{W}}}$.

**Step 1. vectorized form of** $\sum_{q=1}^{m}(y_q-\eta[\phi(a_q)\otimes\phi(x_q)\otimes\widehat{\mu}_{W|a_q,x_q,z_q}])^2$

Let $\widehat{v}:=vec_c(\widehat{\alpha})$, where $\widehat{v}$ is the column-wise vectorization of $\hat{\alpha}$. That is, for $A=\begin{bmatrix}a & b\\ c & d\end{bmatrix}$, the vectorization is $vec_c(A)=\begin{bmatrix}a\\ c\\ b\\ d\end{bmatrix}$. It can be shown that for column-wise vectorization of compatible matrices $K, L$ and $M$, we have:

$$vec(KLM) = (M^T\otimes K)vec_c(L) \tag{48}$$

This equality is known as *Roth's relationship* between vectors and matrices. See:(Macedo & Oliveira, 2013, Eq. 82) for proof of column-wise vectorization. Now, if as a specific case we define:

$$\begin{aligned} M &:= K_{WW}\left[K_{AA}\odot K_{XX}\odot K_{ZZ}+m_1\lambda_w\right]^{-1}\left[K_{A\widetilde{a}_q}\odot K_{X\widetilde{x}_q}\odot K_{Z\widetilde{z}_q}\right]\\ L &:= \alpha^T\\ K &:= K_{\widetilde{a}_q\widetilde{A}}\odot K_{\widetilde{x}_q\widetilde{X}} = \left(K_{\widetilde{A}\widetilde{a}_q}\odot K_{\widetilde{X}\widetilde{x}_q}\right)^T; \end{aligned}$$

In this case, $KLM$ is scalar and $\in\mathbb{R}$ and $vec(KLM)=vec([KLM]^T)$.

Subsequently, we can write:

$$vec(KLM) = (M^T\otimes K)vec_c(L) = vec_c^T(L)(M\otimes K^T). \tag{49}$$

The second equality uses that transposition and conjugate transposition are distributive over the Kronecker product.

By applying (48) to the matrix form of eq. (40), we obtain:

$$\begin{aligned} &\eta[\phi(\widetilde{a}_q)\otimes\phi(\widetilde{x}_q)\otimes\widehat{\mu}_{W|\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q}]\\ =\ & vec^T(\alpha)\left[\left\{K_{WW}\left[K_{AA}\odot K_{XX}\odot K_{ZZ}+m_1\lambda_w\right]^{-1}\left[K_{A\widetilde{a}_q}\odot K_{X\widetilde{x}_q}\odot K_{Z\widetilde{z}_q}\right]\right\}\otimes\left(K_{\widetilde{A}\widetilde{a}_q}\odot K_{\widetilde{X}\widetilde{x}_q}\right)\right] \end{aligned} \tag{50}$$

Notice, that since the column-wise vectorization of a matrix is equal to the row-wise vectorization of its transpose, $vec_c(\alpha^T)=vec(\alpha):=v$.

To derive the vectorized form of (6), (50) can be expanded for $(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)$ for all $q\in\{1,...m_2\}$. Note that in (49), $M$ is the $q$-th column of $C$ defined in (44); and $K^\top$ is the $q$-th column of $K_{\widetilde{A}\widetilde{A}}\odot K_{\widetilde{X}\widetilde{X}}$. To derive the vectorized form of eq. (27),

we expand the results of (49) to all columns of underlying matrices. We introduce operator $\overline{\otimes}$ as a column-wise Kronecker product of matrices[1]. Note that this operator is in fact the *column-wise Khatri–Rao product*.

Finally,

$$\sum_{q=1}^{m_2} \left(y_q - \eta[\phi(\widetilde{a}_q) \otimes \phi(\widetilde{x}_q) \otimes \widehat{\mu}_{W|\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q}]\right)^2 = \|Y - v^T\, C\overline{\otimes}\left[K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right]\|_2^2 = \|Y - v^T D\|_2^2, \tag{51}$$

with $C$ and $D$ defined by (44) and (45).

**Step 2. Expressing $\|\eta\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$ in terms of the vector $v$**

For the regularization term in (6), we use the expression of the norm of $\eta$ in matrix terms as presented in (34):

$$\begin{aligned}
\|\eta\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} =& Tr\left\{\alpha^T K_{WW}\alpha(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}})\right\} \\
&= vec(\alpha)^T vec(K_{WW}\alpha(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}})) \\
&= vec(\alpha)^T\{K_{WW} \otimes (K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}})^T\}vec(\alpha) \\
&= v^T\{K_{WW} \otimes (K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}})\}\, v \qquad (52) \\
&:= v^T E v. \qquad (53)
\end{aligned}$$

Note that the vectorization is row-wise. In the second equality, we used that $Trace(A^T B) = vec(A)^T vec(B)$ for two square matrices $A$ and $B$ of the same size. The third equality is the row-wise expression of *Roth's relationship* between vectors and matrices (see Macedo & Oliveira (2013)). □

### B.5.2 Derivation of the closed form solution for $\widehat{\eta}$

We presented the vectorized form of ERM eq. (6) in Lemma 5. Its minimizer $\widehat{v}$ is the solution to a ridge regression in $\mathbb{R}^{m1m2}$ and its closed-form is easily available through:

$$\widehat{v} = \left\{\left[DD^T + m_2\lambda_2 E\right]^{-1} D\right\} y, \tag{54}$$

with $D$ and $E$ given by (45) and (46), respectively. The solution still requires inversion of $DD^T$, an $m_1m_2 \times m_1m_2$ matrix, however. In the following, we use the Woodbury identity to derive an efficient closed-form solution for eq. (6).

**Lemma 6.** *The closed form solution in eq.* (54) *can be rearranged as:*

$$\begin{aligned}
\widehat{v} &= \left(\Gamma_{(A,X,Z)}\overline{\otimes} I\right)\left(m_2\lambda_2 I + \Sigma\right)^{-1} y \qquad \in \mathbb{R}^{m_1m_2} \qquad (55) \\
\textit{and } \Sigma &= \left[\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW} \Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right)\left(K_{\widetilde{a}_q\widetilde{a}_p} K_{\widetilde{x}_q\widetilde{x}_p}\right)\right]_{m_2\times m_2}, \textit{ for } p,q \in \{1,\dots,m_2\}, \qquad (56)
\end{aligned}$$

where $\Gamma_{(a,x,z)} := \Gamma(a,x,z)$ is defined in (37). Hence, the closed-form solution for $v := vec(\alpha)$ only involves the inversion of an $m_2 \times m_2$ matrix $\Sigma$.

*Proof.* We start by applying Woobudy identity to eq. (54):

$$\begin{aligned}
\widehat{v} &= \left\{\left[DD^T + m_2\lambda_2 E\right]^{-1} D\right\} y \\
&= E^{-1}D\left[m_2\lambda_2 I + D^T E^{-1} D\right]^{-1} y \qquad (57) \\
&= \left(\Gamma_{(A,X,Z)}\overline{\otimes} I\right)\left(m_2\lambda_2 I + \Sigma\right)^{-1} y \qquad \in \mathbb{R}^{m_1m_2} \qquad (58) \\
\text{and } \Sigma &= \left[\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW} \Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right)\left(K_{\widetilde{a}_q\widetilde{a}_p} K_{\widetilde{x}_q\widetilde{x}_p}\right)\right]_{m_2\times m_2}, \text{ for } p,q \in \{1,\dots,m_2\} \qquad (59)
\end{aligned}$$

The final equality, (58), is the outcome of lemma 7. □

[1] $A\overline{\otimes}B = A_i \otimes B_i$ for all $i$s, columns of matrices A and B. This operation is equivalent of Kronecker product of columns and requires matrices A and B to have the same number of columns (but they can have a different number of rows). Note that $\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})_q} = \Gamma_{\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q}$ and $\{K_{\widetilde{A}\widetilde{A}} \odot O_{\widetilde{X}\widetilde{X}}\}_q = K_{\widetilde{A}\widetilde{a}_q} \odot O_{\widetilde{X}\widetilde{x}_q}$, respectively. This operator allows us to express empirical loss in matrix-vector form.

**Lemma 7.** *We may write* $D^T E^{-1} D = \Sigma$, *where* $\Sigma = \left[\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW} \Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right)\left(K_{\widetilde{a}_q\widetilde{a}_p} K_{\widetilde{x}_q\widetilde{x}_p}\right)\right]_{m_2\times m_2}$, *for* $p, q \in \{1,\ldots,m_2\}$.

*Proof.* We first show that: $E^{-1}D = \Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}$

$$
\begin{aligned}
E^{-1}D &= \left(K_{WW} \otimes \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)\right)^{-1} \left(K_{WW}\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)\right) \\
&= \left(K^{-1}_{WW} \otimes \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)^{-1}\right) \left(K_{WW}\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)\right) \\
&= \left[\ldots, \left(K^{-1}_{WW} \otimes \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)^{-1}\right) \left(K_{WW}\Gamma_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} \otimes \left(K_{\widetilde{A}\widetilde{a}_q} \odot K_{\widetilde{X}\widetilde{x}_q}\right)\right), \ldots\right] \\
&= \left[\ldots, \left(K^{-1}_{WW} K_{WW}\Gamma_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)}\right) \otimes \left(\left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)^{-1} \left(K_{\widetilde{A}\widetilde{a}_q} \odot K_{\widetilde{X}\widetilde{x}_q}\right)\right), \ldots\right] \\
&= \left[\ldots, \Gamma_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} \otimes I_q, \ldots\right] = \Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}
\end{aligned}
$$

For the third equality, we expand $\overline{\otimes}$ in terms of the Kronecker product of associated columns of matrices. We then use the property of the Kronecker product $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for compatible $A, B, C,$ and $D$.

In the second step, we replace $E^{-1}D$ with its equivalent derived in step one, and show that: $D^T(E^{-1}D) = D^T\left(\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}\right) = \left[\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW} \Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right)\left(K_{\widetilde{a}_q\widetilde{a}_p} K_{\widetilde{x}_q\widetilde{x}_p}\right)\right]_{m_2\times m_2}$. First,

$$
\begin{aligned}
&D^T\left(\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}\right) = \left\{C^T \overline{\otimes} \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)\right\}\left(\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}\right) \\
&= \left\{\left(K_{WW}\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})}\right)^T \overline{\otimes} \left(K_{\widetilde{A}\widetilde{A}} \odot K_{\widetilde{X}\widetilde{X}}\right)\right\}\left(\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}\right)
\end{aligned}
$$

Next, let's take a closer look at individual elements of the matrix $[.]_{qp}$, the $q$th row of $p$th column.

$$
\begin{aligned}
&\left[D^T\left(\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2\times m_2}\right)\right]_{qp} = \left\{\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW}\right) \otimes \left(K_{\widetilde{A}\widetilde{a}_q} \odot K_{\widetilde{X}\widetilde{x}_q}\right)\right\}\left(\Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)} \otimes I_p\right) && (60) \\
= &\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW} \Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right) \otimes \left(\left(K_{\widetilde{A}\widetilde{a}_q} \odot K_{\widetilde{X}\widetilde{x}_q}\right) I_p\right) && (61) \\
= &\left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)} K_{WW} \Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right)\left(K_{\widetilde{a}_q\widetilde{a}_p} K_{\widetilde{x}_q\widetilde{x}_p}\right) && (62)
\end{aligned}
$$

In (60) we have used the property of Kronecker product $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for compatible $A, B, C,$ and $D$. □

### B.6 Estimating the causal effect

Recall that the causal effect (2) is written $\beta(a) = \int_{\mathcal{X},\mathcal{W}} h(a,x,w) f(x,w) dx dw$ . Since $h \in \mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}$ by Assumption 9, and using the reproducing property, we can write:

$$
\begin{aligned}
\beta(a) &= \int_{\mathcal{X},\mathcal{W}} h(a,x,w)\rho(x,w) dx dw \\
&= \int_{\mathcal{X},\mathcal{W}} \langle h, \phi(a) \otimes \phi(x) \otimes \phi(w)\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \rho(x,w) dx dw \\
&= \langle h, \phi(a) \otimes \int_{\mathcal{X}\mathcal{W}} \phi(x) \otimes \phi(w)\rho(x,w) dx dw\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}. \qquad (63)
\end{aligned}
$$

Consequently, $\widehat{h}$ can be expressed as: $\widehat{h} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \widehat{\alpha}_{ij} \phi(\widetilde{a}_j) \otimes \phi(\widetilde{x}_j) \otimes \phi(w_i)$. We can further replace $\int_{\mathcal{X}\mathcal{W}} \phi(x) \otimes \phi(w)\rho(x,w) dx dw$ by its empirical estimate $\frac{1}{n}\sum_{k=1}^{n} \phi(x_k) \otimes \phi(w_k)$ from the sample $\{(x,w)_k\}_{k=1}^n$. This leads to the

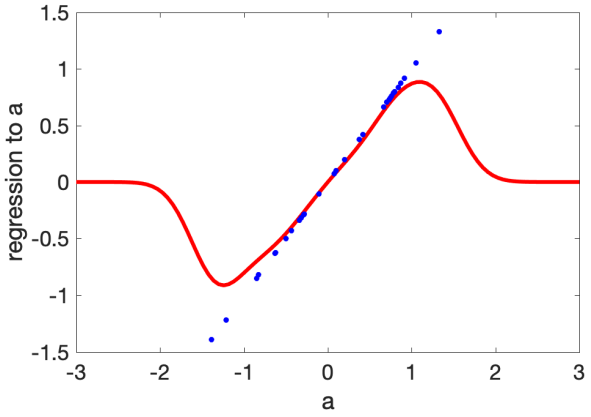


Figure 4: Learning an identity map on a non-compact domain using (Gaussian) kernel ridge regression.

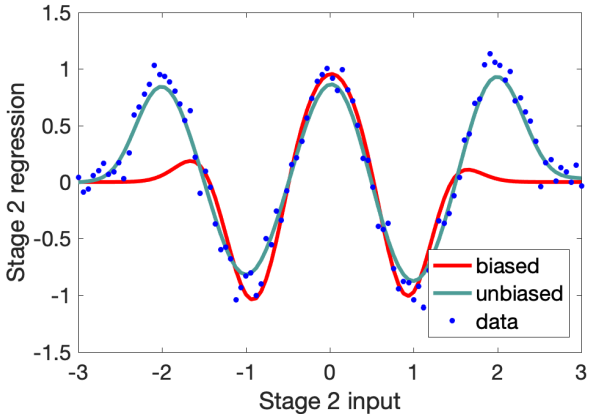


Figure 5: Bias in second stage as a result of using $\hat{\mu}_{A|a}$ in Stage 1 ("biased") vs regressing on $\phi(a)$ ("unbiased").

following estimator of the causal effect:

$$\begin{aligned}\widehat{\beta}(a) &= \langle \widehat{h}, \phi(a) \otimes \int_{\mathcal{XW}} \frac{1}{n}\sum_{k=1}^{n} \phi(x_k)\otimes\phi(w_k)\rangle_{\mathcal{H}_{\mathcal{AXW}}} \\ &= \langle \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \widehat{\alpha}_{ij}\phi(\widetilde{a}_j)\otimes\phi(\widetilde{x}_j)\otimes\phi(w_i), \phi(a)\otimes\frac{1}{n}\sum_{k=1}^{n}\phi(x_k)\otimes\phi(w_k)\rangle_{\mathcal{H}_{\mathcal{AXW}}} \\ &= \frac{1}{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{n}\widehat{\alpha}_{ij}K_{a\widetilde{a}_j}K_{x_k\widetilde{x}_j}K_{w_k w_i}. \end{aligned} \tag{64}$$

### B.7 An alternative two-stage solution, and its shortcomings

As an alternative solution to kernel proximal causal learning, one might consider using the Stage 1 estimate of $\widehat{\mu}_{W,A|A,X,Z}\left(\phi(a)\otimes\phi(x)\otimes\phi(z)\right) := \widehat{\mathbb{E}}\left(\phi(W)\otimes\phi(A)|a,x,z\right)$, obtained by ridge regression, as an input in Stage 2, which would allow an unmodified use of the KIV algorithm (Singh et al., 2019) in the proxy setting. This method has both theoretical and empirical shortcomings, however.

Theoretically, regression from $\phi(a)$ to $\phi(a)$ is, in population limit, the identity mapping $I_{\mathcal{H}_\mathcal{A}}$ from $\mathcal{H}_\mathcal{A}$ to $\mathcal{H}_\mathcal{A}$. This operator is not Hilbert-Schmidt for characteristic RKHSs, and violates the well-posedness assumption for consistency of Stage 1 regression (Singh et al., 2019).

In practice, predicting $\phi(a)$ via ridge regression from $\phi(a)$ introduces bias in the finite sample setting. This is shown in an example in Figures 4 and 5. In a first stage (Figure 4), the identity map is approximated by ridge regression, where the distribution $\rho_\mathcal{A}(a)$ is Gaussian centred at the origin. This distribution is supported on the entire real line, but for finite samples, few points are seen at the tails, and bias is introduced (the function reverts to zero). The impact of this bias will reduce as more training samples are observed (although the identity map will never be learned perfectly, as discussed earlier). This bias affects the second stage. In Figure 5, the distribution of $a$ for the second stage is uniform on the interval $[-3,3]$. This is a *subset* of the stage 1 support of $\rho_\mathcal{A}(a)$, yet due to the limited number of samples from stage 1, bias is nonetheless introduced near the boundaries of that interval. This bias can be more severe as the dimension of $a$ increases. As seen in Figure 5, this bias impacts the second stage, where we compare regression from $\hat{\mu}_{A|a}$ to $y$ (*biased*) with regression from $\phi(a)$ to $y$ (*unbiased*). This bias is avoided in our KPV setting by using the Stage 2 input $\mu_{W|a,z}\otimes\phi(a)$ instead of $\mu_{W,A|a,z}$ (ignoring $x$ for simplicity).

### B.8 Consistency

In this section, we provide consistency results for the KPV approach. For any Hilbert space $\mathcal{F}$, we denote $\mathcal{L}(\mathcal{F})$ the space of bounded linear operators from $\mathcal{F}$ to itself. For any Hilbert space $\mathcal{G}$, we denote by $\mathcal{L}^2(\mathcal{F},\mathcal{G})$ the space of Hilbert-Schmidt operators from $\mathcal{F}$ to $\mathcal{G}$. We denote by $L^2(\mathcal{F},\rho)$ the space of square integrable functions on $\mathcal{F}$ with respect to measure $\rho$.

### B.8.1 THEORETICAL GUARANTEES FOR STAGE 1

The optimal $C_{W|X,A,Z}$ minimizes the expected discrepancy:

$$C_{W|X,A,Z} = \underset{C \in \mathcal{L}^2(\mathcal{H}_{AXZ}, \mathcal{H}_{\mathcal{W}})}{\operatorname{argmin}} E(C), \text{ where } E(C) = \mathbb{E}_{WAXZ} \|\phi(W) - C\phi(A,X,Z)\|^2_{\mathcal{H}_{\mathcal{W}}}$$

We now provide a non-asymptotic consistency result for Stage 1. This directly follows the Stage 1 IV proof of Singh et al. (2019), based in turn on the regression result of Smale & Zhou (2007), and we simply state the main results as they apply in our setting, referencing the relevant theorems from the earlier work as needed.

The problem of learning $C_{W|A,X,Z}$ is transformed into a vector-valued regression, where the search space is the vector-valued RKHS $\mathcal{H}_\Gamma$ of operators mapping $\mathcal{H}_{\mathcal{AXZ}}$ to $\mathcal{H}_{\mathcal{W}}$. A crucial result is that $\mathcal{H}_{\mathcal{AXZ}} \otimes \mathcal{H}_{\mathcal{W}}$ is isomorphic to $\mathcal{L}^2(\mathcal{H}_{\mathcal{AXZ}}, \mathcal{H}_{\mathcal{W}})$. Hence, by choosing the vector-valued kernel $\Gamma$ with feature map : $(a,x,z,w) \mapsto [\phi(a) \otimes \phi(x) \otimes \phi(z) \otimes \phi(w)] := \phi(a) \otimes \phi(x) \otimes \phi(z) \langle \phi(w), \cdot \rangle_{\mathcal{H}_{\mathcal{W}}}$, we have $\mathcal{H}_\Gamma = \mathcal{L}^2(\mathcal{H}_{\mathcal{AXZ}}, \mathcal{H}_{\mathcal{W}})$ and they share the same norm. We denote by $L^2(\mathcal{A} \times \mathcal{X} \times \mathcal{Z}, \rho_{\mathcal{AXZ}})$ the space of square integrable functions from $\mathcal{A} \times \mathcal{X} \times \mathcal{Z}$ to $\mathcal{W}$ with respect to measure $\rho_{\mathcal{AXZ}}$, where $\rho_{\mathcal{AXZ}}$ is the restriction of $\rho$ to $\mathcal{A} \times \mathcal{X} \times \mathcal{Z}$.

**Assumption 12** Suppose that $C_{W|X,A,Z} \in \mathcal{H}_\Gamma$, i.e. $C_{W|X,A,Z} = \operatorname{argmin}_{C \in \mathcal{H}_\Gamma} E(C)$.

**Definition 1** (Kernel Integral operator for Stage 1)**.** *Define the integral operator :*

$$S_1 \colon L^2(\mathcal{A} \times \mathcal{X} \times \mathcal{Z}, \rho_{\mathcal{AXZ}}) \longrightarrow \mathcal{H}_{\mathcal{AXZ}}$$
$$g \longmapsto \int \phi(a,x,z) g(a,x,z) d\rho_{\mathcal{AXZ}}(a,x,z).$$

*The uncentered covariance operator is defined by $T_1 = S_1 \circ S_1^*$, where $S_1^*$ is the adjoint of $S_1$.*

**Assumption 13** Fix $\gamma_1 < \infty$. For given $c_2 \in (1,2]$, define the prior $\mathcal{P}(\gamma_1, c_1)$ as the set of probability distributions $\rho$ on $\mathcal{A} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{W}$ such that a range space assumption is satisfied : $\exists G_1 \in \mathcal{H}_\Gamma$ s.t. $C_{W|A,X,Z} = T_1^{\frac{c_1 - 1}{2}} \circ G_1$ and $\|G_1\|^2_{\mathcal{H}_\Gamma} \leq \gamma_1$.

Our estimator for $C_{W|AXZ}$ is given by ERM (5) based on $\{(a,x,z,w)_i\}_{i=1}^{m_1}$. The following theorem provides the closed-form solution of (5).

**Theorem 4.** *(Singh et al., 2019, Theorem 1) For any $\lambda_1 > 0$, the solution of (5) exists, is unique, and is given by:*

$$\widehat{C}_{W|A,X,Z} = (\boldsymbol{T}_1 + \lambda_1)^{-1} g_1, \text{ where } \boldsymbol{T}_1 = \frac{1}{m_1} \sum_{i=1}^{m_1} \phi(a_i,x_i,z_i) \otimes \phi(a_i,x_i,z_i),$$
$$\text{and } g_1 = \frac{1}{m_1} \sum_{i=1}^{m_1} \phi(a_i,x_i,z_i) \otimes \phi(w_i);$$

*and for any $(a,x,z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$, we have $\widehat{\mu}_{W|a,x,z} = \widehat{C}_{W|A,X,Z} (\phi(a) \otimes \phi(x) \otimes \phi(z))$.*

Under the assumptions provided above, we can now derive a non-asymptotic bound in high probability for the estimated conditional mean embedding, for a well-chosen regularization parameter.

**Theorem 5.** *Suppose Assumptions 5, 7, 12 and 13 hold. Define $\lambda_1$ as:*

$$\lambda_1 = \left( \frac{8\kappa^3(\kappa + \kappa^3 \|C_{W|A,X,Z}\|_{\mathcal{H}_\Gamma}) \ln(2/\delta)}{\sqrt{m_1 \gamma_1 (c_1 - 1)}} \right)^{\frac{2}{c_1 + 1}}$$

*Then, for any $x, a, z \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$ and any $\delta \in (0,1)$, the following holds with probability $1 - \delta$:*

$$\|\widehat{\mu}_{W|a,x,z} - \mu_{W|a,x,z}\|_{\mathcal{H}_{\mathcal{W}}} \leq \kappa^3 r_C(\delta, m_1, c_1) =: \kappa^3 \frac{\sqrt{\gamma_1}(c_1 + 1)}{4^{\frac{1}{c_1 + 1}}} \left( \frac{4\kappa^3(\kappa + \kappa^3 \|C_{W|A,X,Z}\|_{\mathcal{H}_\Gamma}) \ln(2/\delta)}{\sqrt{m_1 \gamma_1}(c_1 - 1)} \right)^{\frac{c_1 - 1}{c_1 + 1}},$$

*where $\widehat{\mu}_{W|a,x,z} = \widehat{C}_{W|A,X,Z} (\phi(a) \otimes \phi(x) \otimes \phi(z))$ and $\widehat{C}_{W|A,X,Z}$ is the solution of (5).*

*Proof.* Under Assumption 5 and 7, $\mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Z}$ are separable (see Lemma 4.33 of Steinwart & Christmann (2008)). Hence, for any $(a, x, z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$, we have : $\|\phi(a) \otimes \phi(x) \otimes \phi(z)\|_{\mathcal{H}_{\mathcal{AXZ}}} = \|\phi(a)\|_{\mathcal{H}_\mathcal{A}} \|\phi(x)\|_{\mathcal{H}_\mathcal{X}} \|\phi(z)\|_{\mathcal{H}_\mathcal{Z}} \leq \kappa^3$ by Assumption 7. Then, we can write:

$$\begin{aligned}\|\widehat{\mu}_{W|a,x,z} - \mu_{W|a,x,z}\|_{\mathcal{H}_\mathcal{W}} &= \|(\widehat{C}_{W|A,X,Z} - C_{W|A,X,Z})\,(\phi(a) \otimes \phi(x) \otimes \phi(z))\,\|_{\mathcal{H}_\mathcal{W}} \\ &\leq \|\widehat{C}_{W|A,X,Z} - C_{W|A,X,Z}\|_{\mathcal{H}_\Gamma} \|\phi(a) \otimes \phi(x) \otimes \phi(z)\|_{\mathcal{H}_{\mathcal{AXZ}}} \\ &\leq \kappa^3 r_C(\delta, m_1, c_1)\end{aligned}$$

where the last inequality results from Singh et al. (2019, Theorem 2). □

### B.8.2 THEORETICAL GUARANTEES FOR STAGE 2

The optimal $\eta$ minimizes the expected discrepancy:

$$\eta_{AXW} = \underset{\eta \in \mathcal{H}_{\mathcal{AXW}}}{\operatorname{argmin}}\ \tilde{R}(\eta), \text{ where } \tilde{R}(\eta) = \mathbb{E}_{AXZY}\left\{Y - \eta[\phi(a,x) \otimes \mu_{W|a,x,z}]\right\}^2.$$

Similarly to Stage 1, the problem of learning $\eta_{AXW}$ is transformed into a ridge regression, where the search space is the RKHS $\mathcal{H}_{\mathcal{AXW}}$ of $\mathcal{Y}$-valued functions ($\mathcal{Y} \subset \mathbb{R}$). We now provide our assumptions to derive non asymptotic results for Stage 2. The approach builds on the Stage 2 proof of Singh et al. (2019), based in turn on (Caponnetto & De Vito, 2007; Szabó et al., 2016), with modifications made to account for the difference in setting, since the input to our Stage 2 differs from the case of instrumental variable regression (see proofs for details).

**Assumption 14** Suppose that $\eta_{AXW} \in \mathcal{H}_{\mathcal{AXW}}$, i.e. $\eta_{AXW} = \operatorname{argmin}_{\eta \in \mathcal{H}_{\mathcal{AXW}}} \tilde{R}(\eta)$.

**Definition 2** (Kernel integral operator for Stage 2)**.** *Define the integral operator :*

$$\begin{aligned}S_2 \colon \mathcal{H}_{\mathcal{AXW}} &\longrightarrow \mathcal{H}_{\mathcal{AXW}} \\ \eta &\longmapsto \int [\mu_{W|a,x,z} \otimes \phi(a,x)]\eta[\phi(a,x) \otimes \mu_{W|a,x,z}] d\rho_{\mathcal{H}_\mathcal{W} \times \mathcal{A} \times \mathcal{X}}(\mu_{W|a,x,z}, a, x).\end{aligned}$$

*The uncentered covariance operator is defined by $T_2 = S_2 \circ S_2^*$, where $S_2^*$ is the adjoint of $S_2$.*

**Assumption 15** Fix $\gamma_2 < \infty$. For given $c_1 \in (1, 2]$, define the prior $\mathcal{P}(\gamma_2, b, c_2)$ as the set of probability distributions $\rho$ on $\mathcal{H}_{\mathcal{AXW}} \times \mathcal{Y}$ such that:

- A range space assumption is satisfied : $\exists G_2 \in \mathcal{H}_{\mathcal{AXW}}$ s.t. $\eta_{AXW} = T_2^{\frac{c_2-1}{2}} \circ G_2$ and $\|G_2\|_{\mathcal{H}_{\mathcal{AXW}}} \leq \gamma_2$
- The eigenvalues $(l_k)_{k \in \mathbb{N}^*}$ of $T_2$ satisfy $\alpha_2 \leq l_k k^{-b_2} \leq \beta_2$ for $b_2 > 1$, $\alpha_2, \beta_2 > 0$.

**Theorem 6.** *Assume Assumptions 5 to 7 and 12 to 15 hold. Assume the assumptions of Theorem 5 hold and define $\lambda_1$ accordingly. Assume also that $m_1, m_2$ are large enough (see Proposition 9) and that $\lambda_2 \leq \|T_2\|_{\mathcal{L}\mathcal{H}_{\mathcal{AXW}}}$. Then, for any $\epsilon, \delta \in (0,1)$, the following holds w.p. $1 - \epsilon - \delta$:*

$$\begin{aligned}\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) \leq r_H(\delta, m_1, c_1, \epsilon, m_2, b_2, c_2) := 5 \Bigg\{ &\frac{4\kappa^{10} c_Y^2}{\lambda_2} r_C(\delta, m_1, c_1)^2 \\ &+ \frac{4\kappa^{10} c_Y^2}{\lambda_2} r_C(\delta, m_1, c_1)^2 . 4 \left( \frac{8 \ln^2(6/\epsilon)}{\lambda_2} \left[ \frac{(c_Y + \|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{AXW}}})^2 (4 + m_2 \lambda_2 (\beta_2^{\frac{1}{b_2}} \frac{\pi/b_2}{sin(\pi/b_2)} \lambda_2^{-\frac{1}{b_2}})}{m_2^2 \lambda_2} \right] + \right. \\ &\left. \frac{2 \ln^2(6/\epsilon)}{\lambda_2} \left[ \frac{4\gamma_2 \lambda_2^{c_2 - 1} + m_2 \gamma_2 \lambda_2^{c_2}}{m_2^2 \lambda_2} \right] + \gamma_2 \lambda_2^{c_2 - 1} + \|\eta_{AXW}\|^2_{\mathcal{H}_{\mathcal{AXW}}} \right) + \gamma_2 \lambda_2^{c_2} \\ &+ 32 \ln^2(6/\epsilon) \left[ \frac{(c_Y + \|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{AXW}}})^2 (4 + m_2 \lambda_2 (\beta_2^{\frac{1}{b_2}} \frac{\pi/b_2}{sin(\pi/b_2)} \lambda_2^{-\frac{1}{b_2}}))}{m_2^2 \lambda_2} \right] + 8 \ln^2(6/\epsilon) \left[ \frac{4\gamma_2 \lambda_2^{c_2 - 1} + m_2 \gamma_2 \lambda_2^{c_2}}{m_2^2 \lambda_2} \right] \Bigg\}.\end{aligned}$$

*Proof.* By Proposition 5, we have:

$$\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) \leq 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda_2) + S_1 + S_2\right].$$

Then, by Proposition 10, w.p. $1 - \frac{\epsilon}{3} - \delta$, we have:

$$S_{-1} \leq \frac{4}{\lambda_2}\kappa^{10} r_C(\delta, m_1, c_1)^2 c_Y^2, \quad S_0 \leq \frac{4}{\lambda_2}\kappa^{10} r_C(\delta, m_1, c_1)^2 \|\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$

where by Proposition 11, w.p. $1 - \frac{2\epsilon}{3}$:

$$\|\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$
$$\leq 4\left(\frac{8\ln^2(6/\epsilon)}{\lambda_2}\left[\frac{(c_Y + \|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}})^2(4 + m_2\lambda_2\mathcal{N}(\lambda_2))}{m_2^2\lambda_2}\right] + \frac{2\ln^2(6/\epsilon)}{\lambda_2}\left[\frac{4\mathcal{B}(\lambda_2) + m_2\mathcal{A}(\lambda_2)}{m_2\lambda_2}\right] + \mathcal{B}(\lambda_2) + \|\eta_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\right).$$

Also, by Proposition 7, w.p. $1 - \frac{2\epsilon}{3}$, we have:

$$S_1 \leq 32\ln^2(6/\epsilon)\left[\frac{(c_Y + \|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}})^2(4 + m_2\lambda_2\mathcal{N}(\lambda_2))}{m_2^2\lambda_2}\right], \quad S_2 \leq 8\ln^2(6/\epsilon)\left[\frac{4\mathcal{B}(\lambda_2) + m_2\mathcal{A}(\lambda_2)}{m_2^2\lambda_2}\right].$$

Finally, by Proposition 6,

$$\mathcal{A}(\lambda_2) \leq \gamma_2\lambda_2^{c_2}, \quad \mathcal{B}(\lambda_2) \leq \gamma_2\lambda_2^{c_2-1}, \quad \mathcal{N}(\lambda_2) \leq \beta_2^{\frac{1}{b_2}}\frac{\pi/b_2}{sin(\pi/b_2)}\lambda_2^{-\frac{1}{b_2}}.$$

Combining all the probabilistic bounds yields the final result. □

**Proof of Theorem 2.**

*Proof.* Ignoring constants in Theorem 6, we have:

$$S_{-1} = O\left(\frac{r_C(\delta, m_1, c_1)^2}{\lambda_2}\right),$$
$$S_0 = O\left(\frac{r_C(\delta, m_1, c_1)^2}{\lambda_2} \cdot \left(\frac{1}{m_2^2\lambda_2^2} + \frac{1}{m_2\lambda_2^{1+1/b_2}} + \frac{1}{m_2^2\lambda_2^{3-c_2}} + \frac{1}{m_2\lambda_2^{2-c_2}} + \lambda_2^{c_2-1} + 1\right)\right)$$
$$\mathcal{A}(\lambda_2) = O(\lambda_2^{c_2}), \quad S_1 = O\left(\frac{1}{m_2^2\lambda_2} + \frac{1}{m_2\lambda_2^{1/b_2}}\right), \quad S_2 = O\left(\frac{1}{m_2^2\lambda_2^{2-c_2}} + \frac{1}{m_2\lambda_2^{1-c_2}}\right).$$

The last term in $S_0$ indicates that $S_0$ dominates $S_{-1}$. Moreover, since $b_2 > 1$ and $c_2 \in (1, 2]$, we have that $\frac{1}{m_2}$ dominates $\frac{1}{m_2\lambda_2^{3-c_2}}$; that $\frac{1}{m_2\lambda_2^{1+1/b_2}}$ dominates $\frac{1}{m_2\lambda_2^{2-c_2}}$; and that 1 dominates $\lambda_2^{c_2-1}$ (since $\lambda_2 \to 0$). For the same reasons, $S_1$ dominates $S_2$.

Hence, we have:

$$\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) = O\left(\frac{r_C(\delta, m_1, c_1)^2}{\lambda_2}\left[\frac{1}{m_2^2\lambda_2^2} + \frac{1}{m_2\lambda_2^{1+1/b_2}} + 1\right] + \lambda_2^{c_2} + \frac{1}{m_2^2\lambda_2} + \frac{1}{m_2\lambda_2^{1/b_2}}\right).$$

By Theorem 5, and by choosing $m_1 = m_2^{\zeta\frac{c_1+1}{c_1-1}}$ as stated in Theorem 2, we have successively:

$$r_C(\delta, m_1, c_1)^2 = O\left(m_1^{-\frac{c_1+1}{c_1-1}}\right) = O(m_2^{-\zeta}),$$

which leads to:

$$\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) = O\left(\frac{1}{m_2^{2+\zeta}\lambda_2^3} + \frac{1}{m_2^{1+\zeta}\lambda_2^{2+1/b_2}} + \frac{1}{m_2^{\zeta}\lambda_2} + \lambda_2^{c_2} + \frac{1}{m_2^2\lambda_2} + \frac{1}{m_2\lambda_2^{1/b_2}}\right).$$

The final result is from Szabó et al. (2016, Theorem 5). □

### B.8.3 PROOF DETAILS FOR THEOREM 6

First introduce $\widetilde{\eta}_{AXW}$ as the minimizer of the empirical risk of stage 2, when plugging the true $\mu_{W|a,x,z}$ (instead of its estimate from Stage 1):

$$\widetilde{\eta}_{AXW} = \underset{\eta \in \mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}{\operatorname{argmin}} \widetilde{L}(\eta), \text{ where } \widetilde{L}(\eta) = \frac{1}{m_2}\sum_{j=1}^{m_2}\left(\widetilde{y}_j - \eta[\phi(\widetilde{a}_j, \widetilde{x}_j) \otimes \mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}]\right)^2 + \lambda_2\|\eta\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}. \tag{65}$$

Similarly to $\widehat{\eta}_{AXW}$, it has a closed form solution given below (see Grunewalder et al. (2012, Section D.1)).

**Theorem 7.** *For any $\lambda_2 > 0$, the solutions of (65), exists, is unique, and is given by:*

$$\widetilde{\eta}_{AXW} = (\boldsymbol{T}_2 + \lambda_2)^{-1} g_2, \text{ where } \boldsymbol{T}_2 = \frac{1}{m_2}\sum_{j=1}^{m_2}\left[\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} \otimes \phi(\widetilde{a}_j, \widetilde{x}_j)\right] \otimes \left[\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} \otimes \phi(\widetilde{a}_j, \widetilde{x}_j)\right]$$

$$\text{and } g_2 = \frac{1}{m_2}\sum_{j=1}^{m_2}\left[\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} \otimes \phi(\widetilde{a}_j, \widetilde{x}_j)\right]\widetilde{y}_j.$$

Define also $\eta^{\lambda_2}_{AXW}$ as the minimizer of the population version of (65):

$$\eta^{\lambda_2}_{AXW} = \underset{\eta \in \mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}{\operatorname{argmin}} L^{\lambda_2}(\eta), \text{ where } L^{\lambda_2}(\eta) = \mathbb{E}_{AXYZ}\left\{Y - \eta[\phi(A, X) \otimes \mu_{W|a,x,z}]\right\}^2 + \lambda_2\|\eta\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}. \tag{66}$$

The excess risk for the KPV estimator can be decomposed in five terms as stated in the following proposition.

**Proposition 5.** *The excess risk of the Stage 2 estimator can be bounded by five terms:*

$$\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) \leq 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda_2) + S_1 + S_2\right]$$

*where*

$$S_{-1} = \|\sqrt{T_2} \circ (\widehat{\boldsymbol{T}}_2 + \lambda_2)^{-1}(\widehat{g}_2 - g_2)\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}, \quad S_0 = \|\sqrt{T_2} \circ (\widehat{\boldsymbol{T}}_2 + \lambda_2)^{-1} \circ (\boldsymbol{T}_2 - \widehat{\boldsymbol{T}}_2)\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$

$$S_1 = \|\sqrt{T_2} \circ (\boldsymbol{T}_2 + \lambda_2)^{-1}(g_2 - \boldsymbol{T}_2\eta_{AXW})\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}, \quad S_2 = \|\sqrt{T_2} \circ (\boldsymbol{T}_2 + \lambda_2)^{-1} \circ (T_2 - \boldsymbol{T}_2)(\eta^{\lambda_2}_{AXW} - \eta_{AXW})\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$

*and the residual* $\mathcal{A}(\lambda_2) = \|\sqrt{T_2}(\eta^{\lambda_2}_{AXW} - \eta_{AXW})\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$.

*Proof.* The excess risk can be decomposed as:

$$\begin{aligned}\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) &= \|\sqrt{T_2}(\widehat{\eta}_{AXW} - \eta_{AXW})\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \\ &= \|\sqrt{T_2}\left[(\widehat{\eta}_{AXW} - \widetilde{\eta}_{AXW}) + (\widetilde{\eta}_{AXW} - \eta^{\lambda_2}_{AXW}) + (\eta^{\lambda_2}_{AXW} - \eta_{AXW})\right]\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\end{aligned} \tag{67}$$

Using the operator identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ and Theorem 1, the first term in (67) can be bounded by $5(S_{-1} + S_0)$, the second one by $5(S_1 + S_2)$ and the last one by $5\mathcal{A}(\lambda_2)$ (see Szabó et al. (2015, Section A.1.8)). The factor 5 comes from the inequality $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$. □

The first two terms $S_{-1}, S_0$ characterize the estimation error due to Stage 1; the middle term $\mathcal{A}(\lambda_2)$ characterizes the regularization bias; while the two last terms $S_1, S_2$ characterize the estimation error from Stage 2. The goal is now to bound each term of Proposition 5 separately. For the three last terms from Stage 2, we can benefit from the minimax rates and results for ridge regression (Caponnetto & De Vito, 2007), see Propositions 6 and 7. Stage 1 requires intermediate results (Propositions 8 to 10).

We firstly have the following bounds that characterize the relation between $\eta^{\lambda_2}_{AXW}$ and $\eta_{AXW}$.

**Proposition 6.** *Suppose Assumption 15 holds, which means that $\rho \in \mathcal{P}(\gamma_2, b_2, c_2)$ and that the eigenvalues $(l_k)_{k\in\mathbb{N}^*}$ of $T_2$ satisfy $\alpha_2 \leq l_k k^{-b_2} \leq \beta_2$. Then, the residual $\mathcal{A}(\lambda_2)$, the reconstruction error $\mathcal{B}(\lambda_2)$, and the effective dimension $\mathcal{N}(\lambda_2)$ are defined and bounded as follows:*

$$\mathcal{A}(\lambda_2) = \|\sqrt{T_2}(\eta^{\lambda_2}_{AXW} - \eta_{AXW})\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \leq \gamma_2\lambda_2^{c_2}, \qquad \mathcal{B}(\lambda_2) = \|\eta^{\lambda_2}_{AXW} - \eta_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \leq \gamma_2\lambda_2^{c_2-1},$$

$$\mathcal{N}(\lambda_2) = Tr\left[(T_2 + \lambda_2)^{-1} \circ T_2\right] \leq \beta_2^{\frac{1}{b_2}}\frac{\pi/b_2}{sin(\pi/b_2)}\lambda_2^{-\frac{1}{b_2}}.$$

The bounds on $\mathcal{A}(\lambda_2)$, $\mathcal{B}(\lambda_2)$ follow from Caponnetto & De Vito (2007, Proposition 3), while the bound on $\mathcal{N}(\lambda_2)$ follows from Sutherland (2017). The residual $\mathcal{A}(\lambda_2)$ and reconstruction error $\mathcal{B}(\lambda_2)$, which depend on $\rho$, control the complexity of $\eta_{AXW}$. The effective dimension $\mathcal{N}(\lambda_2)$ measures the complexity of the hypothesis space $\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}$ with respect to $\rho_{\mathcal{H}_\mathcal{W}\times\mathcal{A}\times\mathcal{X}}$.

**Proposition 7.** *(Caponnetto & De Vito, 2007, Step 2 and 3 of Theorem 4) Assume Assumption 6 and Assumption 14 hold. Assume also that $\lambda_2 \leq \|T_2\|_{\mathcal{L}\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$ and $m_2 \geq \frac{2C_\epsilon\mathcal{N}(\lambda_2)}{\lambda_2}$. Then, we can bound $S_1$ and $S_2$ from Proposition 5 as follows w.p. $1-2\epsilon/3$:*

$$S_1 \leq 32\ln^2(6/\epsilon)\left[\frac{(c_Y+\|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}})^2(4+m_2\lambda_2\mathcal{N}(\lambda_2))}{m_2^2\lambda_2}\right], \quad S_2 \leq 8\ln^2(6/\epsilon)\left[\frac{4\mathcal{B}(\lambda_2)+m_2\mathcal{A}(\lambda_2)}{m_2^2\lambda_2}\right].$$

The following bounds are obtained easily by using the bounds from Theorem 5 on the difference between the estimated conditional mean embeddings of Stage 1 and the true one.

**Proposition 8.** *Assume the assumptions of Theorem 5 hold and define $\lambda_1$ accordingly. Suppose also that Assumptions 6 and 14 hold. Then, w.p. $1-\delta$ :*

$$\|\widehat{g}_2 - g_2\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \leq \kappa^{10} r_C(\delta, m_1, c_1)^2 c_Y^2, \quad \text{and} \quad \|\boldsymbol{T}_2 - \widehat{\boldsymbol{T}}_2\|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq 4\kappa^{10} r_C(\delta, m_1, c_1)^2. \tag{68}$$

*Proof.* Using Assumption 6, Theorem 5, and $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$, we have :

$$\begin{aligned}
\|\widehat{g}_2 - g_2\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} &\leq \frac{1}{m_2}\sum_{j=1}^{m_2} \| \left[(\widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} - \mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}) \otimes \phi(\widetilde{a}_j,\widetilde{x}_j)\right]\widetilde{y}_j\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \\
&\leq \frac{1}{m_2}\sum_{j=1}^{m_2} \|(\widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} - \mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j})\|^2_{\mathcal{H}_\mathcal{W}} \|\phi(\widetilde{a}_j,\widetilde{x}_j)\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}}} \widetilde{y}_j^2 \leq \kappa^{10} r_C(\delta, m_1, c_1)^2 c_Y^2.
\end{aligned}$$

On the other hand, using $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$ and the identity $(a+b)^2 \leq 2a^2+2b^2$, we have:

$$\begin{aligned}
&\|\boldsymbol{T}_2 - \widehat{\boldsymbol{T}}_2\|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \\
&\leq \frac{2}{m_2}\sum_{j=1}^{m_2} \| \left[(\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} - \widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}) \otimes \phi(\widetilde{a}_j,\widetilde{x}_j)\right] \otimes \left[\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} \otimes \phi(\widetilde{a}_j,\widetilde{x}_j)\right] \|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \\
&+ \frac{2}{m_2}\sum_{j=1}^{m_2} \| \left[\widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} \otimes \phi(\widetilde{a}_j,\widetilde{x}_j)\right] \otimes \left[(\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} - \widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}) \otimes \phi(\widetilde{a}_j,\widetilde{x}_j)\right] \|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \\
&\leq \frac{2}{m_2}\sum_{j=1}^{m_2} \|\widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} - \mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}\|^2_{\mathcal{H}_\mathcal{W}} \|\mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}\|^2_{\mathcal{H}_\mathcal{W}} \|\phi(\widetilde{a}_j,\widetilde{x}_j)\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}}} \\
&+ \frac{2}{m_2}\sum_{j=1}^{m_2} \|\widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j} - \mu_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}\|^2_{\mathcal{H}_\mathcal{W}} \|\widehat{\mu}_{W|\widetilde{a}_j,\widetilde{x}_j,\widetilde{z}_j}\|^2_{\mathcal{H}_\mathcal{W}} \|\phi(\widetilde{a}_j,\widetilde{x}_j)\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}}} \\
&\leq 4\kappa^{10} r_C(\delta, m_1, c_1)^2.
\end{aligned}$$

□

**Proposition 9.** *Assume the assumptions of Theorem 5 hold and define $\lambda_1$ accordingly. Let $C_\epsilon = 96\ln^2(6/\epsilon)$. Suppose also that Assumptions 6 and 14 hold. Finally, assume $\lambda_2 \leq \|T_2\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})}$ and that :*

$$m_2 \geq \frac{2C_\epsilon\mathcal{N}(\lambda_2)}{\lambda_2}, \qquad m_1 \geq \bar{m}(\delta, c_1)), := \left[\frac{8\kappa\sqrt{\gamma_1}(c_1+1)}{4^{\frac{1}{c_1+1}}\lambda_2}\right]^{2\frac{c_1+1}{c_1-1}} \left(\frac{4\kappa^3(\kappa+\kappa^3\|C_{W|A,X,Z}\|_{\mathcal{H}_\Gamma})\ln(2/\delta)}{\sqrt{\gamma_1}(c_1-1)}\right)^2.$$

*Then, w.p. $1-\frac{\epsilon}{3}-\delta$, we have:*

$$\|\sqrt{T_2}\circ(\widehat{\boldsymbol{T}}_{\boldsymbol{2}}+\lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \frac{2}{\sqrt{\lambda_2}}.$$

*Proof.* We follow the proof of Singh et al. (2019, Proposition 39). Using the Neumann series of $I - (T_2 - \widehat{\boldsymbol{T}}_{\boldsymbol{2}})(T_2 + \lambda_2)^{-1}$, we have:

$$\|\sqrt{T_2} \circ (\widehat{\boldsymbol{T}}_{\boldsymbol{2}} + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \|\sqrt{T_2} \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \sum_{k=0}^{\infty} \|(T_2 - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|^k_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})}.$$

We first deal with the first term on the r.h.s. Observe that by definition of the operator norm,

$$\|\sqrt{T_2} \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} = \sup_{l \in (l_k)_{k \in \mathbb{N}^*}} \frac{\sqrt{l}}{l + \lambda_2} \leq \frac{1}{2\sqrt{\lambda_2}},$$

where the last inequality results from arithmetic-geometric mean inequality ($\sqrt{l\lambda_2} \leq (l + \lambda_2)/2$). We now deal with the second term on the r.h.s. First, we apply the triangle inequality :

$$\|(T_2 - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \|(T_2 - \boldsymbol{T}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} + \|(\boldsymbol{T}_{\boldsymbol{2}} - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})}.$$

Since $\|(T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq 1/\lambda_2$, by Proposition 9 the second term is easily bounded w.p. $1 - \delta$ as :

$$\|(\boldsymbol{T}_{\boldsymbol{2}} - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \frac{\|\boldsymbol{T}_{\boldsymbol{2}} - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})}}{\lambda_2} \leq \frac{\kappa^5 r_C(\delta, m_1, c_1)}{\lambda_2}.$$

For a fixed $\lambda_2$, $m_1$ can be chosen so that $\kappa^5 r_C(\delta, m_1, c_1)/\lambda_2 \leq 1/4$, which legitimates the use of the Neumann series at the beginning of the proof. This is actually given by setting $m_1 \geq \bar{m}(\delta, c_1)$. By Caponnetto & De Vito (2007, Step 2.1, Theorem 4), the first term is bounded with probability $1 - \frac{\epsilon}{3}$ by:

$$\|(T_2 - \boldsymbol{T}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \frac{1}{2}.$$

for $m_2 \geq \frac{2C_\epsilon \mathcal{N}(\lambda_2)}{\lambda_2}$. Hence, we can conclude that for $m_1 \geq \bar{m}(\delta, c_1)$ and $m_2 \geq \frac{2C_\epsilon \mathcal{N}(\lambda_2)}{\lambda_2}$, we have w.p. $1 - \frac{\epsilon}{3} - \delta$:

$$\|(T_2 - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}) \circ (T_2 + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \frac{1}{2} + \frac{1}{4} = \frac{3}{4} \Longrightarrow \|\sqrt{T_2} \circ (\widehat{\boldsymbol{T}}_{\boldsymbol{2}} + \lambda_2)^{-1}\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \leq \frac{1}{2\sqrt{\lambda_2}} \frac{1}{1 - \frac{3}{4}} = \frac{2}{\sqrt{\lambda_2}}. \quad \square$$

We now bound each term separately.

**Proposition 10.** *Assume the conditions of Propositions 8 and 9 hold. We can bound $S_{-1}$ and $S_0$ from Proposition 5 w.p. $1 - \frac{\epsilon}{3} - \delta$ as follows:*

$$S_{-1} \leq \frac{4}{\lambda_2} \kappa^{10} r_C(\delta, m_1, c_1)^2 c_Y^2, \quad S_0 \leq \frac{4}{\lambda_2} \kappa^{10} r_C(\delta, m_1, c_1)^2 \|\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}.$$

*Proof.* Using Proposition 8 and Proposition 9, we have:

$$S_{-1} \leq \|\sqrt{T_2} \circ (\widehat{\boldsymbol{T}}_{\boldsymbol{2}} + \lambda_2)^{-1}\|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \|\widehat{g}_2 - g_2\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \leq \frac{4}{\lambda_2} \kappa^{10} r_C(\delta, m_1, c_1)^2 c_Y^2,$$

and similarly we have:

$$S_0 \leq \|\sqrt{T_2} \circ (\widehat{\boldsymbol{T}}_{\boldsymbol{2}} + \lambda_2)^{-1}\|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \|\boldsymbol{T}_{\boldsymbol{2}} - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}\|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})} \|\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \leq \frac{4}{\lambda_2} \kappa^{10} r_C(\delta, m_1, c_1)^2 \|\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}. \quad \square$$

**Proposition 11.** *Let $C_\epsilon = 96 \ln^2(6/\epsilon)$ and suppose that $m_2 \geq \frac{2C_\epsilon \mathcal{N}(\lambda_2)}{\lambda_2}$ and that $\lambda_2 \leq \|T_2\|_{\mathcal{L}(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}})}$. Then, w.p. $1 - 2\epsilon/3$*

$$\|\widetilde{\eta}_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}$$
$$\leq 4 \left( \frac{8 \ln^2(6/\epsilon)}{\lambda_2} \left[ \frac{(c_Y + \|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}})^2 (4 + m_2 \lambda_2 \mathcal{N}(\lambda_2))}{m_2^2 \lambda_2} \right] + \frac{2 \ln^2(6/\epsilon)}{\lambda_2} \left[ \frac{4\mathcal{B}(\lambda_2) + m_2 \mathcal{A}(\lambda_2)}{m_2^2 \lambda_2} \right] + \mathcal{B}(\lambda_2) + \|\eta_{AXW}\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \right).$$

*Proof.* Using the triangle inequality, we have:

$$\begin{aligned}\|\widetilde{\eta}_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}} &\leq \|\widetilde{\eta}_{AXW} - \eta^{\lambda_2}_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}} + \|\eta^{\lambda_2}_{AXW} - \eta_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}} + \|\eta_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}} \\ &= \|\widetilde{\eta}_{AXW} - \eta^{\lambda_2}_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}} + \sqrt{\mathcal{B}(\lambda_2)} + \|\eta_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}} \\ &\leq \frac{1}{\sqrt{\lambda_2}}(S_1 + S_2) + \sqrt{\mathcal{B}(\lambda_2)} + \|\eta_{AXW}\|_{\mathcal{A}\mathcal{X}\mathcal{W}}\end{aligned}$$

Using $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$ and Proposition 7, we get the final result. □

### B.9 Proof of Proposition 1

Let $\widehat{\mu}_{XW} = \frac{1}{n_t}\sum_{i=1}^{n_t}[\phi(x_i)\otimes\phi(w_i)]$ and $\mu_{XW} = \mathbb{E}_{XW}[\phi(X)\otimes\phi(W)]$. By Tolstikhin et al. (2017, Proposition 1), we have w.p. $1-\delta$:

$$\|\widehat{\mu}_{XW} - \mu_{XW}\|_{\mathcal{H}_{\mathcal{X}\mathcal{W}}} \leq r_\mu(n_t,\delta) \leq \frac{4\kappa^2\ln(2/\delta)}{n_t} := r_\mu(n_t,\delta).$$

Moreover, by Theorem 6, we have w.p. $1-\epsilon-\delta$:

$$\|\widehat{\eta}_{AXW} - \eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \leq r_H(\delta, m_1, c_1, \epsilon, m_2, b_2, c_2).$$

We use the following decomposition for the causal effect :

$$\begin{aligned}\widehat{\beta}(a) - \beta(a) &= \widehat{\eta}_{AXW}[\widehat{\mu}_{XW}\otimes\phi(a)] - \eta_{AXW}[\mu_{XW}\otimes\phi(a)] \\ &= \widehat{\eta}_{AXW}[(\widehat{\mu}_{XW} - \mu_{XW}\otimes\phi(a)] + (\widehat{\eta}_{AXW} - \eta_{AXW})[\mu_{XW}\otimes\phi(a)] \\ &= (\widehat{\eta}_{AXW} - \eta_{AXW})[(\widehat{\mu}_{XW} - \mu_{XW})\otimes\phi(a)] + \eta_{AXW}[(\widehat{\mu}_{XW} - \mathbb{E}_{XW}[\phi(X)\otimes\phi(W)])\otimes\phi(a)] \\ &+ (\widehat{\eta}_{AXW} - \eta_{AXW})[\mu_{XW}\otimes\phi(a)].\end{aligned}$$

Therefore, w.p. $1-\epsilon-\delta$, by Theorem 2, $\|\widehat{\eta}_{AXW} - \eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} = 0(m^{-\alpha})$ with $\alpha \in \left\{\frac{\zeta c_2}{c_2+1}, \frac{b_2c_2}{b_2c_2+1}\right\}$ and :

$$\begin{aligned}|\widehat{\beta}(a) - \beta(a)| &\leq \|\widehat{\eta}_{AXW} - \eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\|\widehat{\mu}_{XW} - \mu_{XW}\|_{\mathcal{H}_{\mathcal{X}\mathcal{W}}}\|\phi(a)\|_{\mathcal{H}_{\mathcal{A}}} + \|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\|\widehat{\mu}_{XW} - \mu_{XW}\|_{\mathcal{H}_{\mathcal{X}\mathcal{W}}}\|\phi(a)\|_{\mathcal{H}_{\mathcal{A}}} \\ &+ \|\widehat{\eta}_{AXW} - \eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}\|\mu_{XW}\otimes\phi(a)\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \\ &\leq \kappa.r_H(\delta, m_1, c_1, \epsilon, m_2, b_2, c_2)r_\mu(n_t,\delta) + \kappa\|\eta_{AXW}\|_{\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}}r_\mu(n_t,\delta) + \kappa^3 r_H(\delta, m_1, c_1, \epsilon, m_2, b_2, c_2) \\ &= O(n_t^{-\frac{1}{2}} + m^{-\alpha}).\end{aligned}$$

## C Proxy Maximum Moment Restriction

In this section, we propose a novel approach to solve the proximal causal learning using the maximum moment restriction (MMR) framework (Muandet et al., 2020a). It is based on that proposed by Zhang et al. (2020) for the IV setting. On the other hand, we adapt it to the proxy setting, with a novel interpretation for $h$. This is inspired by Miao et al. (2018) and Tchetgen Tchetgen et al. (2020), but in their formulations $h$ is defined to be the solution of an ill-posed inverse problem, whereas we view $h$ as a regression function for $y$, which is more interpretable, as we will detail below.

### C.1 Maximum Moment Restriction for Proxy Setting

**Notations.** (i) Let $\mathcal{X}$ denote a measurable space. (ii) Let $X$ denote a random variable taking values in $\mathcal{X}$.

Notice that $\{Y, A, X, W\}$ are random variables under the generating process governed by Figure 1. Let $h \in \Omega(\mathcal{A}\times\mathcal{W}\times\mathcal{X})$ be a measurable function on $\mathcal{A}\times\mathcal{W}\times\mathcal{X}$. Therefore, $h(A, W, X)$ is a function of random variables, which is a random variable itself.

**Proof of Lemma 1.** The proposed method is based on Lemma 1, which shows that any function $h \in \Omega(\mathcal{A}\times\mathcal{W}\times\mathcal{X})$ that is the solution to Equation (1) must also satisfy the conditional moment restriction (CMR), and vice versa.

*Proof.* Let $\varepsilon$ be a random variable representing the residual of $h(A, W, X)$ with respect to $Y$:

$$\varepsilon := Y - h(A, W, X). \tag{69}$$

Suppose that $h$ is the solution to (1). Then, taking the conditional expectation of (69) conditioned on $A, Z, X$ yields

$$\begin{aligned}\mathbb{E}[\varepsilon|A,Z,X] &= \mathbb{E}[Y|A,Z,X] - \mathbb{E}[h(A,W,X)|A,X,Z] \\ &= \mathbb{E}[Y|A,Z,X] - \int_{\mathcal{W}} h(A,w,X) f(w|A,Z,X) dw \\ &= 0.\end{aligned}$$

In the last term of the second equality, we take expectation over $W$ because it is the only variable not being held fixed by conditioning. The last equality holds because $h$ is the solution to (1) by definition. □

Note that we have derived the condition typical in additive noise instrumental variable (IV) models (Hartford et al., 2017; Dikkala et al., 2020; Bennett et al., 2019; Zhang et al., 2020; Muandet et al., 2020b). A more general term for this type of conditions is called *conditional moment restrictions* (CMR) (Newey, 1993). This interpretation allows us to approach the problem of learning $h$ from a different perspective. That is to say, we look for $h$ for which the conditional moment restriction is zero. This contrasts with the typical two-stage approach of learning $h$, for which the objective is to find $h$ such that $\mathbb{E}_{AXZY}[(Y - \mathbb{E}_W[h(A,W,X)|A,Z,X])^2]$ is minimized.

**Connection to IV.** Typical formulation of IV models assumes the following structural model:

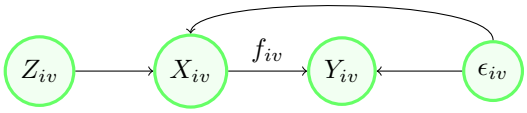


Figure 6: DAG of an instrumental variable model

where in particular $Z_{iv} \perp\!\!\!\perp \epsilon_{iv}$. Additionally, an additive noise model for generating $Y_{iv}$ is typically assumed, and the noise is assumed to have zero mean. In mathematical terms, these amount to

$$Y_{iv} = f_{iv}(X_{iv}) + \epsilon_{iv}, \qquad \mathbb{E}[\epsilon_{iv}] = 0, \qquad \mathbb{E}[\epsilon_{iv}|Z_{iv}] = 0 \quad a.s.$$

**Remark 5.** *We make two comparisons between the IV setting and our proxy setting.*

1. *In the IV setting, $Y_{iv} = f_{iv}(X_{iv}) + \epsilon_{iv}$ is proposed as the structural equation for $Y_{iv}$; in contrast, in our proxy setting $Y = h(A,X,W) + \epsilon$ is not a structural equation, as it does not remain invariant under interventions. For readers unfamiliar with the concept of structural equations, we refer to Pearl (2000). We thus offer the interpretation that $Y = h(A,W,X) + \epsilon$ is a regression equation whose noise term has mean zero when conditioning on $A, Z, X$.*
2. *In the IV setting, authors make the assumption of additive noise models for the structural equation of $Y_{iv}$, and $f_{iv}$ is then the causal effect of $X_{iv}$ on $Y_{iv}$, i.e. $f_{iv}(X_{iv}) = \mathbb{E}[Y_{iv}|do(X_{iv})]$. In our setting, the solution $h$ of the integral equation (1) directly gives us the causal effect, hence no additive noise assumption is made and our approach is entirely nonparametric.*
3. *Lemma 1 establishes the connection between a class of problems that can be formulated in terms of an integral equation like (1) and those that satisfy the CMR. Hence, we believe this result can be applied more broadly to problems that share similar structure to our setting.*

**Proof of Lemma 2**

*Proof.* Since $g \in \mathcal{H}_{\mathcal{AZX}}$, we may write $g(A,Z,X) = \langle g, k((A,Z,X),\cdot)\rangle$, thus we have

$$\begin{aligned}R_k(h) &= \sup_{g\in\mathcal{H}_{\mathcal{AZX}}, \|g\|\leq 1} (\mathbb{E}[(Y - h(A,W,X))\langle g, k((A,Z,X),\cdot)\rangle])^2 \\ &= \sup_{g\in\mathcal{H}_{\mathcal{AZX}}, \|g\|\leq 1} (\mathbb{E}[\langle g, (Y - h(A,W,X))k((A,Z,X),\cdot)\rangle])^2 \\ &= \sup_{g\in\mathcal{H}_{\mathcal{AZX}}, \|g\|\leq 1} (\langle g, \mathbb{E}[(Y - h(A,W,X))k((A,Z,X),\cdot)]\rangle)^2 \\ &= \|\mathbb{E}[(Y - h(A,W,X))k((A,Z,X),\cdot)]\|^2_{\mathcal{H}_{\mathcal{AZX}}}.\end{aligned}$$

The second equality is due to linearity of an inner product, and we remark that it still holds despite $h$ and $g$ sharing variables $A$ and $X$, because $(Y - h(A, W, X)) \in \mathbb{R}$ as opposed to $\mathcal{H}_{\mathcal{A}\mathcal{W}\mathcal{X}}$. The last equality is due to the fact that $\mathcal{H}_{\mathcal{A}\mathcal{Z}\mathcal{X}}$ is a vector space, and $\mathbb{E}[(Y - h(A, W, X))k((A, Z, X), \cdot)] \in \mathcal{H}_{\mathcal{A}\mathcal{Z}\mathcal{X}}$ by assumption.

Then,

$$\begin{aligned} R_k(h) &= \langle \mathbb{E}[(Y - h(A, W, X))k((A, Z, X), \cdot), \mathbb{E}[(Y - h(A, W, X))k((A, Z, X), \cdot)\rangle \\ &= \mathbb{E}[\langle (Y - h(A, W, X))k((A, Z, X), \cdot), (Y' - h(A', W', X'))k((A', Z', X'), \cdot)\rangle] \\ &= \mathbb{E}[(Y - h(A, W, X))(Y' - h(A', W', X')k((A, Z, X), (A', Z', X')))], \end{aligned}$$

as required. □

### C.2 Analytical Solution for PMMR

Suppose further that $h$ also lies in an RKHS $\mathcal{H}_{\mathcal{A}\mathcal{Z}\mathcal{X}}$ endowed with the kernel function $l$. Then, we can use the representer theorem (Schölkopf et al., 2001) to derive a close-form solution for $h$. We note that the risk functional $R_k$ is different from standard least squares risk since it involves independent data samples as well as the kernel function $k$. Nevertheless, the empirical risk still applies to data samples $\{y_i, a_i, w_i, x_i, z_i\}_{i=1}^n$, so the representer theorem still apply on RKHS features $\{l((a_i, w_i, x_i), \cdot)\}_{i=1}^n$. This is to say, that by the representer theorem,

$$\hat{h}(a, w, x) = \sum_{i=1}^{n} \alpha_i l((a_i, w_i, x_i), (a, w, x)),$$

for some $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. Hence, we may rewrite the optimization problem as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\operatorname{argmin}} \, (\mathbf{y} - L\boldsymbol{\alpha})^\top W (\mathbf{y} - L\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top L \boldsymbol{\alpha} \tag{70}$$

where $L_{ij} = l((a_i, w_i, x_i), (a_j, w_j, x_j))$ and $W_{ij} = k((a_i, z_i, x_i), (a_j, z_j, x_j))$. The solution to (70) can be found by solving the first-order stationary condition, resulting in the closed-form expression:

$$\hat{\boldsymbol{\alpha}} = (LWL + \lambda L)^{-1} LW\mathbf{y}.$$

It can be shown that $\mathcal{H}_{\mathcal{X}_1 \times \cdots \times \mathcal{X}_m}$ is isometrically isomorphic to $\mathcal{H}_{\mathcal{X}_1} \otimes \cdots \otimes \mathcal{H}_{\mathcal{X}_m}$. In the latter, the kernel of the outer-product RKHS can be decomposed into the product of the kernels of the children RKHSes:

$$k(\mathbf{x}, \mathbf{x}') = k_1(x_1, x_1')k_2(x_2, x_2') \cdots k_m(x_m, x_m').$$

Hence, we may use an alternative closed-form formulation of $h$ with the product kernels

$$\hat{h}(a, w, x) = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i l_{\mathcal{A}}(a_i, a) l_{\mathcal{W}}(w_i, w) l_{\mathcal{X}}(x_i, x)$$

#### C.2.1 Applying the representation theorem to PMMR

First, we quote the representation theorem.

**Theorem 8.** *Consider a positive-definite real-valued kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ with a corresponding reproducing kernel Hilbert space $H_k$. Let there be given*

- *a training sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$,*
- *a strictly increasing real-valued function $g : [0, \infty) \to \mathbb{R}$, and*
- *an arbitrary error function $E : \left(\mathcal{X} \times \mathbb{R}^2\right)^n \to \mathbb{R} \cup \{\infty\}$,*

*which together define the following regularized empirical risk functional on* $H_k$ *:*

$$f \mapsto E\left(\left(x_1, y_1, f\left(x_1\right)\right), \ldots, \left(x_n, y_n, f\left(x_n\right)\right)\right) + g(\|f\|)$$

*Then, any minimizer of the empirical risk*

$$f^* = \operatorname{argmin}_{f \in H_k} \left\{E\left(\left(x_1, y_1, f\left(x_1\right)\right), \ldots, \left(x_n, y_n, f\left(x_n\right)\right)\right) + g(\|f\|)\right\}, \quad (*)$$

*admits a representation of the form:*

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k\left(\cdot, x_i\right)$$

*where* $\alpha_i \in \mathbb{R}$ *for all* $1 \leq i \leq n$*.*

In our case, we have

$$\begin{aligned} E : (\mathcal{X} \times \mathbb{R}^2)^n &\rightarrow \mathbb{R} \cup \{\infty\} \\ \{((a_i, w_i, x_i, z_i), h(a_i, w_i, x_i), y_i)\}_{i=1}^n &\mapsto \sum_{i,j=1}^{n} \frac{\left(y_i - h\left(a_i, w_i, x_i\right)\right)\left(y_j - h\left(a_j, w_j, x_j\right)\right) k\left(\left(a_i, z_i, x_i\right), \left(a_j, z_j, x_j\right)\right)}{n^2} \end{aligned}$$

so the representer theorem gives us $h(a, w, x) = \sum_{i=1}^{n} \alpha_i l\left(\left(a_i, w_i, x_i\right), (a, w, x)\right)$.

### C.3 PMMR Algorithm

PMMR algorithm to estimate $h$ and derive causal effect is summarized below:

**Algorithm 2** PMMR Algorithm

**input :** 1. Train data $\{z_i^t, w_i^t, a_i^t, y_i^t, x_i^t\}_{i=1}^n$. 2. Kernel functions $l$ for $\mathcal{H}_{\mathcal{AWX}}$ and $k$ for $\mathcal{H}_{\mathcal{AZX}}$ with bandwidths $\sigma_k$ and $\sigma_l$ respectively. 3. Regularisation parameter $\lambda$. 4. Nyström approximation size $M$.

**output :** $\hat{h}(a, w, x)$

```
/* Write x for the matrix containing x_i in the ith row.                    */
```

For all $1 \leq i \leq n$, $1 \leq j \leq n$, $K_{ij} \leftarrow k((a_i, z_i, x_i), (a_j, z_j, x_j))$

For all $1 \leq i \leq n$, $1 \leq j \leq n$, $L_{ij} \leftarrow l((a_i, w_i, x_i), (a_j, w_j, x_j))$

Do Nyström approximation for $K/n^2$, decomposing into $K/n^2 = \tilde{U}\tilde{V}\tilde{U}^T$

$\hat{\alpha} \leftarrow \lambda^{-1}[I - \tilde{U}(\lambda^{-1}\tilde{U}^T L\tilde{U} + \tilde{V}^{-1})^{-1}\tilde{U}^T\lambda^{-1}L]\tilde{U}\tilde{V}\tilde{U}^T\mathbf{y}$

$\hat{h}(a, w, x) \leftarrow l((a, w, x), (\mathbf{a}^t, \mathbf{w}^t, \mathbf{x}^t))$

### C.4 Consistency and Convergence Rates

In this section, we provide a consistency result of the causal estimate as well as the convergence rate of the PMMR solution. For this, we will need the consistency result of the kernel mean embedding.

**Lemma 8** (Proposition A.1, Tolstikhin et al. (2017))**.** *In the following, the authors present a general result whose special cases establishes the convergence rate of* $n^{-1/2}$ *for* $\|\mu_k(P_n) - \mu_k(P)\|_{\mathcal{F}}$ *when* $\mathcal{F} = \mathcal{H}_k$ *and* $\mathcal{F} = L^2(\mathcal{R}^d)$*. They denote* $P_n(X) := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$*.*

*Let* $(X_i)_{i=1}^n$ *be random samples drawn* $i.i.d.$ *from* $P$ *defined on a separable topological space* $\mathcal{X}$*. Suppose* $g : \mathcal{X} \rightarrow H$ *is continuous and*

$$\sup_{x \in \mathcal{X}} \|g(x)\|_H^2 < C_k < \infty \quad (71)$$

*where* $H$ *is a separable Hilbert space of real-valued functions. Then, for any* $0 < \delta \leq 1$ *with probability at least* $1-\delta$ *we have*

$$\left\| \int_{\mathcal{X}} g(x) dP_n(x) - \int_{\mathcal{X}} g(x) dP(x) \right\|_H \leq \sqrt{\frac{C_k}{n}} + \sqrt{\frac{2C_k \log(1/\delta)}{n}} = r(n, \delta). \quad (72)$$

### C.4.1 PMMR CONSISTENCY

**Definition 3.** *For clarity, we define the following variables.*

- $h_0(a, x, w)$ *is a solution to (1).*
- $\hat{h}_n$ *is the solution of a learning algorithm with sample size* $n$.
- $\beta(a) := \mathbb{E}_{WX}[h_0(a, W, X)] = \mathbb{E}[Y|do(a)]$.
- $\hat{\beta}_n(a) := \mathbb{E}_{WX}[\hat{h}_n(a, W, X)]$.
- $\hat{\beta}_n^m(a) := \frac{1}{m}\sum_{i=1}^m \hat{h}_n(a, w_i, x_i)$ *with* $\{w_i, x_i\}_{i=1}^m \sim_{i.i.d.} \mathcal{P}_{\mathcal{WX}}$.
- $\hat{\beta}(a)$*: where* $m$ *and* $n$ *are clear in context, we abuse the notation to write* $\hat{\beta}(a)$ *to denote the estimator for* $\beta(a)$ *from our algorithm.*
- $\mu_X$ *denotes the kernel mean embedding of a random variable* $X$.
- $\hat{\mu}_X^m$ *denotes the empirical estimate of* $\mu_X$*, given by* $\hat{\mu}_X^m = \frac{1}{m}\sum_{i=1}^m[\phi(x_i \sim \mathcal{P}_X)]$*. Where clear from context, we omit the superscript* $m$*, and just use* $\hat{\mu}_X$ *to denote finite-sample estimator of* $\mu_X$.

**Lemma 9** (Causal consistency)**.** *If* $\hat{h}_n \xrightarrow{P} h_0$*, then* $\hat{\beta}_n^m(a) \xrightarrow{P} \beta(a)$ *as* $m, n \to \infty$.

*Proof.* For brevity, in the proof that follows, we write $\mu := \mu_{XW}$ and $\hat{\mu}^m := \hat{\mu}_{XW}^m$.

Since $\beta(a) = \langle h, \mu(\mathcal{P}_{\mathcal{WX}}) \otimes \phi(a)\rangle$ and $\hat{\beta}_n^m(a) = \langle \hat{h}_n, \hat{\mu}^m(\mathcal{P}_{\mathcal{WX}}) \otimes \phi(a)\rangle$, we can verify that

$$\hat{\beta}_n^m(a) - \beta(a) = \langle \hat{h}_n, \phi(a) \otimes \hat{\mu}^m\rangle - \langle h_0, \phi(a) \otimes \mu\rangle \tag{73}$$

$$= \langle \hat{h}_n, \phi(a) \otimes \hat{\mu}^m\rangle - \langle \hat{h}_n, \phi(a) \otimes \mu\rangle + \langle \hat{h}_n, \phi(a) \otimes \mu\rangle - \langle h_0, \phi(a) \otimes \mu\rangle \tag{74}$$

$$= \langle \hat{h}_n, \phi(a) \otimes (\hat{\mu}^m - \mu)\rangle + \langle \hat{h}_n - h_0, \phi(a) \otimes \mu\rangle \tag{75}$$

Thus, by the Cauchy-Schwartz inequality, we have for all $a$,

$$|\hat{\beta}_n^m(a) - \beta(a)| \leq \|\hat{h}\|_{\mathcal{HAWX}} \|\phi(a)\|_{\mathcal{HA}} \|\hat{\mu}^m - \mu\|_{\mathcal{HAW}} + \|\hat{h}_n - h_0\|_{\mathcal{HAWX}} \|\phi(a)\|_{\mathcal{HA}} \|\mu\|_{\mathcal{HAW}} \tag{76}$$

From Lemma 8, by setting $g$ to be the feature map on $\mathcal{A} \times \mathcal{W}$, we have

$$\|\hat{\mu}^m - \mu\| \leq \sqrt{\frac{C_k}{m}} + \sqrt{\frac{2C_k \log(1/\delta)}{m}} =: r(m, \delta) \tag{77}$$

with probability at least $1 - \delta$.

Moreover, we know that $\hat{h}_n \xrightarrow{P} h_0$. This is to say, for any $\epsilon, \delta$, $\exists N$ s.t.

$$\|\hat{h}_n - h_0\|_{\mathcal{HAWX}} \leq \epsilon, \qquad \forall n \geq N \tag{78}$$

with probability at least $1 - \delta$.

Therefore, reflecting on (76) we observe that $\|\mu\|_{\mathcal{HAW}}$ is bounded because we assume bounded kernels, $\|\hat{h}\|_{\mathcal{HAWX}} \xrightarrow{P} \|h_0\|_{\mathcal{HAWX}}$ by assumption, and $\|h_0\|_{\mathcal{HAWX}}$ is constant, $\|\hat{h}_n - h_0\|_{\mathcal{HAWX}}$ and $\|\hat{\mu}^m - \mu\|_{\mathcal{HAW}}$ uniformly converge to zero in probability, which we have just shown.

Therefore, $\sup_{a\in\mathcal{A}} \{\hat{\beta}_n^m(a) - \beta(a)\} \xrightarrow{P} 0$. □

**Lemma 10.** *Suppose* $R_k$ *has at least one minimiser in* $\mathcal{H}_{\mathcal{AWX}}$ *and* $\mathcal{P}_{\mathcal{AWX}}$ *is a finite Borel measure with full support, i.e.,* $\text{supp}[\mathcal{P}_{\mathcal{AWX}}] = \mathcal{A} \times \mathcal{W} \times \mathcal{X}$*. Then,* $R_k$ *has a unique minimiser in* $\mathcal{H}_{\mathcal{AWX}}$ *if and only if the following condition holds:*

$(*)$ $\forall g \in \mathcal{H}_{\mathcal{AWX}}$, $\mathbb{E}_{AWX}[g(A, W, X)|A, Z, X] = 0$ $\mathcal{P}_{\mathcal{AZX}}$*-almost surely if and only if* $g(a, w, x) = 0$ $\mathcal{P}_{\mathcal{AWX}}-$*almost surely.*

*Proof.* To prove that $R_k$ has a unique minimiser in $\mathcal{H}_{\mathcal{AXW}}$, we need i) A minimiser to $R_k$ exists in $\mathcal{H}_{\mathcal{AWX}}$ ii) It is unique. By Assumption 4 and Miao et al. (2018, Appendix, Conditions (v)-(vii)), a minimiser exists in $\mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})$ - we further require $R_k$ has a minimiser in $\mathcal{H}_{\mathcal{AWX}}$ by assumption. We still need to show uniqueness.

( $\implies$ ) Suppose that there exist two different functions $h_1$ and $h_2$ that minimise $R_k$. Then, it follows from Zhang et al. (2020, Theorem 1) that $\mathbb{E}[Y - h_1|A, Z, X] = \mathbb{E}[Y - h_2|A, Z, X] = 0$ $\mathcal{P}_{\mathcal{AZX}}-$almost surely. This means that $\mathbb{E}_{AWX}[h_1(A, W, X) - h_2(A, W, X)|A, Z, X] = 0$ $\mathcal{P}_{\mathcal{AZX}}-$almost surely. Suppose $(*)$ is true, we must have $g(a, w, x) = h_1(a, w, x) - h_2(a, w, x) = 0$ $\mathcal{P}_{\mathcal{AWX}}-$almost surely. As a result, $g$ is the zero function in $\mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})$ and for any other functions $f \in \mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})$, $\langle g, f\rangle_{\mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})} = 0$.

Now, we describe briefly the integral operator representation of our kernel function $l$ and consequently a representation of the RKHS inner product. A more detailed discussion can be found in, e.g., Sejdinovic & Gretton (2014).

**Integral operator of kernel on $\mathcal{H}_{\mathcal{AWX}}$.** Let $l : (\mathcal{A}\times\mathcal{W}\times\mathcal{X})^2 \to \mathbb{R}$ be the kernel function on $\mathcal{A}\times\mathcal{W}\times\mathcal{X}$. We define an operator $S_l : \mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}}) \to \mathcal{C}(\mathcal{A}\times\mathcal{W}\times\mathcal{X})$, where $\mathcal{C}(\mathcal{A}\times\mathcal{W}\times\mathcal{X})$ is the space of continuous functions on $\mathcal{A}\times\mathcal{W}\times\mathcal{X}$, as

$$(S_l f)((a, w, x)) = \int l((a, w, x), (a', w', x'))f((a', w', x'))\, d\mathcal{P}_{\mathcal{AWX}}((a', w', x')), \quad f \in \mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}}) \tag{79}$$

where $S_l$ can be shown to be well-defined (Sejdinovic & Gretton, 2014), and $T_l = I_l \circ S_l$ its composition with the inclusion $I_l : \mathcal{C}(\mathcal{A}\times\mathcal{W}\times\mathcal{X}) \hookrightarrow \mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})$. $T_l$ is said to be the integral operator of kernel $l$.

It can be shown that the symmetry of $l$ implies the integral operator is self-adjoint; the positive definiteness of $l$ implies that $T_l$ is a positive operator, i.e., all eigenvalues are non-negative; continuity of $l$ imples $T_l$ is compact by the Arzela-Ascoli theorem. Then, by the Spectral theorem (Sejdinovic & Gretton, 2014, Theorem 49), any compact, self-adjoint operator can be diagonalised in an appropriate orthonormal basis.

**Relating the RKHS norm with the $\mathcal{L}^2-$norm.** Further supposing that $\mathcal{P}_{\mathcal{AWX}}$ has full support, i.e., $\text{supp}[\mathcal{P}_{\mathcal{AWX}}] = \mathcal{A}\times\mathcal{W}\times\mathcal{X}$, then Mercer's theorem says that for a continuous kernel $l$ on a compact metric space with a finite Borel measure of full support, we can decompose the kernel function $l$ using its at most countable set $J$ of *strictly positive* eigenvalues $\{\lambda_j\}_{j\in J}$ and eigenfunctions $\{e_j\}_{j\in J}$.

$$l((a, w, x), (a', w', x')) = \sum_{j\in J} \lambda_j e_j((a, w, x))e_j((a', w', x')) \tag{80}$$

where the convergence is uniform on $(\mathcal{A}\times\mathcal{W}\times\mathcal{X})^2$ and absolute on each $(a, w, x), (a', w', x') \in \mathcal{A}\times\mathcal{W}\times\mathcal{X}$. See Sejdinovic & Gretton (2014, Section 6.2) for further details.

Then, we may construct the RKHS $\mathcal{H}_{\mathcal{AWX}}$ based on the integral operator $T_l$ and its associated eigenfunctions $\{e_j\}_{j\in J}$, which depend on the underlying measure $\mathcal{P}_{\mathcal{AZX}}$, as

$$\mathcal{H}_{\mathcal{AWX}} = \left\{ f = \sum_{j\in J} a_j e_j \qquad \left\{ \frac{a_j}{\sqrt{\lambda_j}} \right\} \in l^2(J) \right\} \tag{81}$$

with an inner product $\langle \sum_{j\in J} a_j e_j, \sum_{j\in J} b_j e_j\rangle_{\mathcal{H}_{\mathcal{AWX}}} = \sum_{j\in J} \frac{a_j b_j}{\lambda_j}$. Note that $\left\{a_j/\sqrt{\lambda_j}\right\} \in l^2(J)$ implies that $f \in \mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})$. Thus, $a_j = \langle f, e_j\rangle_{\mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})}$.

Now, recall that $g(a, w, x) = h_1(a, w, x) - h_2(a, w, x)$ is a zero function in $\mathcal{L}^2(\mathcal{A}\times\mathcal{W}\times\mathcal{X}, \mathcal{P}_{\mathcal{AWX}})$, which also means that $\|g\|_{\mathcal{H}_{\mathcal{AWX}}} = \sqrt{\langle g, g\rangle_{\mathcal{H}_{\mathcal{AWX}}}} = 0$. Therefore, $h_1(a, w, x) = h_2(a, w, x)$ for all $(a, w, x) \in \mathcal{A}\times\mathcal{W}\times\mathcal{X}$ as norm convergence in RKHS implies pointwise convergence (Steinwart & Christmann, 2008, pp. 119). By contradiction, the minimizer of $R_k$ must be unique.

( $\impliedby$ ) Suppose $(*)$ does not hold, i.e., $(A, Z, X)$ is not complete for $(A, W, X)$, then there exists $g \in \mathcal{H}_{\mathcal{AWX}}$ such that $g \neq 0$ and $\mathbb{E}_{AWX}[g(A, W, X)|A, Z, X] = 0$, $\mathcal{P}_{\mathcal{AZX}}-$almost surely. Then, for any minimizer $h$ of $R_k$ (if it exists), $h + Cg$ for some constant $C$ is also a minimizer, so it cannot be unique. □

**Theorem 9** (Causal consistency of PMMR). *Assume $\mathcal{H}$ is a real-RKHS, $k : (\mathcal{A} \times \mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}$ is bounded, $\Omega(h)$ is convex, $\lambda \xrightarrow{P} 0$. Moreover, assume $Z$ is complete for $W$, i.e., for all $g \in \mathcal{L}^2[\mathcal{P}_\mathcal{W}]$, $\mathbb{E}[g(W)|Z] = 0$, $\mathcal{P}_\mathcal{Z}-$almost surely if and only if $g(W) = 0$, $\mathcal{P}_\mathcal{W}-$almost surely. Then, $\hat{h}_n \xrightarrow{P} h_0$.*

*Proof.* Given $\Omega(h)$ is convex in $h$, we prove consistency based on Newey & McFadden (1994, Theorem 2.7), which requires (i) $R_k(h)$ is uniquely minimized at $h_0$; (ii) $\hat{R}_V(h) + \lambda\Omega(h)$ is convex; (iii) $\hat{R}_V(h) + \lambda\Omega(h) \xrightarrow{P} R_k(h)$ for all $h \in \mathcal{H}$.

Since $\mathcal{H}$ is a real-RKHS, which is a vector space, it is convex because for any $x, y \in \mathcal{H}$, $a \in [0, 1]$, $ax + (1 - a)y \in \mathcal{H}$ by closure of vector spaces. Since $\mathcal{H}$ is convex, $R_k$ is convex (Zhang et al., 2020, Theorem 5). By assumption, $A, Z, X$ is complete for $W$. Then, by Lemma 10, $R_k$ is minimized at $h_0$. Since $\mathcal{H}$ is open, $h_0$ is in the interior of $\mathcal{H}$.

Since $\hat{R}_V(h) = \left\| \frac{1}{n} \sum_{i=1}^n (y_i - h(a_i, w_i, x_i))k((a_i, z_i, x_i), \cdot) \right\|^2_{\mathcal{H}_k}$, by the law of large numbers, we have that $\frac{1}{n} \sum_{i=1}^n (y_i - h(a_i, w_i, x_i))k((a_i, z_i, x_i), \cdot) \xrightarrow{P} \mathbb{E}[(Y - h(A, W, X))k((A, Z, X), \cdot)]$. Then $\hat{R}_V(h) \xrightarrow{P} R_k(h)$ for all $h \in \mathcal{H}$ by the Continuous Mapping Theorem (Mann & Wald, 1943) since $\| \cdot \|_{\mathcal{H}_k}$ is continuous. As $\lambda \xrightarrow{P} 0$, $\hat{R}_V(h) + \lambda\Omega(h) \xrightarrow{P} R_k(h)$ by Slutsky's Theorem (Van der Vaart, 2000, Lemma 2.8). Since $\Omega(h)$ is convex, $\hat{R}_V(h) + \lambda\Omega(h)$ is convex since addition preserves convexity. Thus, by Newey & McFadden (1994, Theorem 2.7), $\hat{h}_n \xrightarrow{P} h_0$. □

**Corollary 1.** *Assume $\mathcal{H}$ is a real-RKHS, $k : (\mathcal{A} \times \mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}$ is bounded, $\Omega(h)$ is convex, and $\lambda \xrightarrow{P} 0$. Moreover, assume $(A, Z, X)$ is complete for $W$, then the causal effect estimate $\hat{\beta}^m_n \xrightarrow{P} 0$ as $m, n \to \infty$.*

*Proof.* By Theorem 9, the conditions guarantee that $\hat{h}_n \xrightarrow{P} h_0$. Then, by Lemma 9 $\hat{\beta}^m_n(A) \xrightarrow{P} \beta(A)$. □

### C.4.2 PMMR Convergence Rate

To provide the convergence rate of PMMR, we will first provide an alternative interpretation of PMMR as a linear ill-posed inverse problem in the RKHS (Nashed & Wahba, 1974; Carrasco et al., 2007). Let $\phi(a, x, w) := k((a, x, w), \cdot)$ and $\varphi(a, x, z) := k((a, x, z), \cdot)$ be the canonical feature maps. Then, the unregularized PMMR objective can be expressed as

$$\begin{aligned} R_k(h) &= \|\mathbb{E}[(Y - h(A, X, W))\varphi(A, X, Z)]\|^2_{\mathcal{H}_{AXZ}} \\ &= \|\mathbb{E}[Y\varphi(A, X, Z)] - \mathbb{E}[h(A, X, W)\varphi(A, X, Z)]\|^2_{\mathcal{H}_{AXZ}} \\ &= \|g - Th\|^2_{\mathcal{H}_{AXZ}}, \end{aligned}$$

where

$$g := \int Y\varphi(A, X, Z)\, d\rho(A, X, Y, Z), \quad Th := \int h(A, X, W)\varphi(A, X, Z)\, d\rho(A, X, W, Z). \tag{82}$$

Here $\rho(A, X, Y, Z)$ and $\rho(A, X, W, Z)$ are the restrictions of $\rho(A, X, W, Y, Z)$ to $\mathcal{A} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ and $\mathcal{A} \times \mathcal{X} \times \mathcal{W} \times \mathcal{Z}$, respectively. By Assumptions 6 and 7, $g \in \mathcal{H}_{AXZ}$ and $T$ is a bounded linear operator from $\mathcal{H}_{AXW}$ to $\mathcal{H}_{AXZ}$. Let $T^* : \mathcal{H}_{AXZ} \to \mathcal{H}_{AXW}$ be an adjoint operator of $T$ such that $\langle Tu, v\rangle_{\mathcal{H}_{AXZ}} = \langle u, T^*v\rangle_{\mathcal{H}_{AXW}}$ for all $u \in \mathcal{H}_{AXW}$ and $v \in \mathcal{H}_{AXZ}$.

Based on the above formulation, we can rewrite the PMMR regularized objective and its empirical estimate as follow:

$$R_\lambda(h) = \|g - Th\|^2_{\mathcal{H}_{AXZ}} + \lambda\|h\|^2_{\mathcal{H}_{AXW}}, \qquad \widehat{R}_\lambda(h) = \|\hat{g} - \widehat{T}h\|^2_{\mathcal{H}_{AXZ}} + \lambda\|h\|^2_{\mathcal{H}_{AXW}}, \tag{83}$$

where $\hat{g}$ and $\widehat{T}$ are the empirical estimates of $g$ and $T$ based on the i.i.d. sample $(a_i, x_i, w_i, y_i, z_i)_{i=1}^n$ from $\rho(A, X, W, Y, Z)$:

$$\hat{g} := \frac{1}{n} \sum_{i=1}^n y_i\varphi(a_i, x_i, z_i), \quad \widehat{T}h := \frac{1}{n} \sum_{i=1}^n h(a_i, x_i, w_i)\varphi(a_i, x_i, z_i). \tag{84}$$

Likewise, we denote by $\widehat{T}^*$ an adjoint operator of $\widehat{T}$, i.e., for $f \in \mathcal{H}_{AXZ}$,

$$T^*f := \int f(A, X, Z)\phi(A, X, W)\, d\rho(A, X, W, Z), \quad \widehat{T}^*f := \frac{1}{n} \sum_{i=1}^n f(a_i, x_i, z_i)\phi(a_i, x_i, w_i). \tag{85}$$

**Cross-covariance operator.** We can view the operator $T$ as an element of the product RKHS $\mathcal{H}_{AXW} \otimes \mathcal{H}_{AXZ}$, i.e., for $h \in \mathcal{H}_{AXW}$,

$$\begin{aligned} Th &= \int h(A,X,W)\varphi(A,X,Z)\, d\rho(A,X,W,Z) \\ &= \int \langle h, \phi(A,X,W)\rangle_{\mathcal{H}_{AXW}} \varphi(A,X,Z)\, d\rho(A,X,W,Z) \\ &= \int [\phi(A,X,W) \otimes \varphi(A,X,Z)]\, h\, d\rho(A,X,W,Z) \\ &= \left[\int \phi(A,X,W) \otimes \varphi(A,X,Z)\, d\rho(A,X,W,Z)\right] h. \end{aligned}$$

Thus, $T = \mathbb{E}[\phi(A,X,W) \otimes \varphi(A,X,Z)] \in \mathcal{H}_{AXW} \otimes \mathcal{H}_{AXZ}$ and is a (uncentered) *cross-covariance* operator mapping from $\mathcal{H}_{AXW}$ to $\mathcal{H}_{AXZ}$ (Baker, 1973; Fukumizu et al., 2004). Likewise, $T^* = \mathbb{E}[\varphi(A,X,Z) \otimes \phi(A,X,W)] \in \mathcal{H}_{AXZ} \otimes \mathcal{H}_{AXW}$. The cross-covariance operator $T$ is Hilbert-Schmidt, and $\|T\| \leq \|T\|_{\text{HS}} = \|T\|_{\mathcal{H}_{AXW} \otimes \mathcal{H}_{AXZ}}$ where $\|\cdot\|_{\text{HS}}$ denotes a Hilbert-Schmidt norm (Fukumizu et al., 2006, Lemma 3). The latter equality holds because the space of Hilbert-Schmidt operators $\text{HS}(\mathcal{H}_1, \mathcal{H}_2)$ forms Hilbert space which are isomorphic to the product space $\mathcal{H}_1 \otimes \mathcal{H}_2$ given by the product kernel.

**PMMR solutions.** Based on (83), we can define the PMMR solutions in the population limit and in the finite sample regime respectively as

$$h_\lambda := \arg\min_{h \in \mathcal{H}_{AXW}} R_\lambda(h) = (T^*T + \lambda I)^{-1} T^* g \tag{86}$$

$$\hat{h}_\lambda := \arg\min_{h \in \mathcal{H}_{AXW}} \widehat{R}_\lambda(h) = (\widehat{T}^*\widehat{T} + \lambda I)^{-1} \widehat{T}^* \hat{g} \tag{87}$$

The solution (86) is obtained by noting that $R_\lambda(h) = \langle h, T^*Th + \lambda h - 2T^*g\rangle_{\mathcal{H}_{AXW}} + \|g\|^2_{\mathcal{H}_{AXZ}}$ whose Frechet derivative is zero only if $(T^*T + \lambda I)h = T^*g$. The solution in (87) can be obtained in a similar way. Let $h_0$ be the solution that uniquely minimizes the unregularized risk $R(h)$. Then, we can decompose the estimation bias into two parts:

$$\hat{h}_\lambda - h_0 = (\hat{h}_\lambda - h_\lambda) + (h_\lambda - h_0). \tag{88}$$

The first part $\hat{h}_\lambda - h_\lambda$ corresponds to an estimation error of the regularized solution $h_\lambda$, whereas the second part $h_\lambda - h_0$ is the regularization bias. Hence, we can obtain the convergence rate of $\hat{h}_\lambda$ by first characterizing the rates of the regularization bias and estimation error separately, and then choosing the regularization parameter $\lambda$ such that both rates coincide.

### Characterizing the Regularization Bias

To control the regularization bias, we impose a regularity condition on the true unknown $h_0$. Following Carrasco et al. (2007), we assume that $h_0$ belong to a regularity space $H_\gamma = (T^*T)^\gamma$ for some positive $\gamma$. The following is a restatement of Carrasco et al. (2007, Def. 3.4); see, also Smale & Zhou (2007) for a similar condition.

**Definition 4** ($\gamma$-regularity space)**.** *The $\gamma$-regularity space of the compact operator $T$ is defined for all $\gamma > 0$, as the RKHS associated with $(T^*T)^\gamma$. That is,*

$$H_\gamma = \left\{ h \in \mathcal{N}(T)^\perp \quad \textit{such that} \quad \sum_{j=1}^{\infty} \frac{\langle h, \phi_j\rangle}{\alpha_j^{2\gamma}} < \infty \right\} \tag{89}$$

*with the inner product*

$$\langle f, g\rangle_\gamma = \sum_{j=1}^{\infty} \frac{\langle f, \phi_j\rangle\langle g, \phi_j\rangle}{\alpha_j^{2\gamma}} \tag{90}$$

*for $f, g \in H_\gamma$.*

In what follows, we will make the following assumption on $h_0$.

**Assumption 16** $h_0 \in H_\gamma$ for $\gamma \in (0, 2]$.

**Proposition 12** (Regularization bias). *Let $T : \mathcal{H}_{AXW} \to \mathcal{H}_{AXZ}$ be an injective compact operator. Then, if Assumption 16 holds and $h_\lambda$ is defined by (86), we have*

$$\|h_\lambda - h_0\|^2_{\mathcal{H}_{AXW}} = \mathcal{O}(\lambda^{\min(\gamma,2)}). \tag{91}$$

*Proof.* Carrasco et al. (2007, Proposition 3.12) □

### Characterizing the Estimation Error

**Proposition 13** (Estimation error). *Let $h_\lambda = (T^*T + \lambda I)^{-1}T^*g$ be the regularized solution given by (86) and $\hat{h}_\lambda = (\widehat{T}^*\widehat{T} + \lambda I)^{-1}\widehat{T}^*\hat{g}$, then*

$$\|\hat{h}_\lambda - h_\lambda\|_{\mathcal{H}_{AXW}} \leq d(\lambda)\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\| + d(\lambda)\|\widehat{T}^*\widehat{T} - T^*T\|\,\|h_0 - h_\lambda\|_{\mathcal{H}_{AXW}}.$$

*where $d(\lambda) := \|\widehat{\Gamma}_\lambda\| = \|(\widehat{T}^*\widehat{T} + \lambda I)^{-1}\|$.*

*Proof.* To simplify the notation, we will use $\Gamma_\lambda := (T^*T + \lambda I)^{-1}$ and $\widehat{\Gamma}_\lambda := (\widehat{T}^*\widehat{T} + \lambda I)^{-1}$ throughout the proof. First, we have

$$\hat{h}_\lambda - h_\lambda = \widehat{\Gamma}_\lambda\widehat{T}^*\hat{g} - \Gamma_\lambda T^*g = \widehat{\Gamma}_\lambda\widehat{T}^*\left(\hat{g} - \widehat{T}h_0\right) + \underbrace{\widehat{\Gamma}_\lambda\widehat{T}^*\widehat{T}h_0 - \Gamma_\lambda T^*Th_0}_{(\star)}. \tag{92}$$

Then, we can write $(\star)$ as

$$\begin{aligned}
\widehat{\Gamma}_\lambda\widehat{T}^*\widehat{T}h_0 - \Gamma_\lambda T^*Th_0 &= \widehat{\Gamma}_\lambda(\widehat{T}^*\widehat{T} - T^*T)h_0 + \widehat{\Gamma}_\lambda T^*Th_0 + \Gamma_\lambda T^*Th_0 \\
&= \widehat{\Gamma}_\lambda(\widehat{T}^*\widehat{T} - T^*T)h_0 + (\widehat{\Gamma}_\lambda - \Gamma_\lambda)T^*Th_0 \\
&\overset{(a)}{=} \widehat{\Gamma}_\lambda(\widehat{T}^*\widehat{T} - T^*T)h_0 + \widehat{\Gamma}_\lambda(T^*T - \widehat{T}^*\widehat{T})\Gamma_\lambda T^*Th_0 \\
&\overset{(b)}{=} \widehat{\Gamma}_\lambda(\widehat{T}^*\widehat{T} - T^*T)h_0 + \widehat{\Gamma}_\lambda(T^*T - \widehat{T}^*\widehat{T})h_\lambda \\
&= \widehat{\Gamma}_\lambda(\widehat{T}^*\widehat{T} - T^*T)(h_0 - h_\lambda),
\end{aligned} \tag{93}$$

where we applied the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ to $\widehat{\Gamma}_\lambda - \Gamma_\lambda$ to get $(a)$, and $(b)$ holds because $h_\lambda = \Gamma_\lambda T^*Th_0$. Combining (92) and (93) yields

$$\hat{h}_\lambda - h_\lambda = \widehat{\Gamma}_\lambda\widehat{T}^*(\hat{g} - \widehat{T}h_0) + \widehat{\Gamma}_\lambda(\widehat{T}^*\widehat{T} - T^*T)(h_0 - h_\lambda).$$

Consequently, we have

$$\|\hat{h}_\lambda - h_\lambda\|_{\mathcal{H}_{AXW}} \leq d(\lambda)\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\| + d(\lambda)\|\widehat{T}^*\widehat{T} - T^*T\|\,\|h_0 - h_\lambda\|_{\mathcal{H}_{AXW}},$$

where $d(\lambda) := \|\widehat{\Gamma}_\lambda\| = \|(\widehat{T}^*\widehat{T} + \lambda I)^{-1}\|$ as required. □

By Proposition 12, Proposition 13, and (88), we can see that the rate of convergence of the estimation bias $\|\hat{h}_\lambda - h_0\|$ depends on the following quantities: (i) A sequence of regularization parameters $\lambda$ which will govern the rate of convergence of the regularization bias $\|h_\lambda - h_0\|$. (ii) The rate of convergence to infinity of $d(\lambda)$. (iii) The rates of convergence of $\|\widehat{T}^*\widehat{T} - T^*T\|$ and $\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\|$ which are governed by the estimation of $T$ and $g$. In the next section, we provide the rates for these intermediate quantities.

### Rates of Intermediate Quantities

Since we will deal with random variables taking values in Hilbert spaces, we need the following concentration inequality.

**Lemma 11** (Bennett inequality in Hilbert space). *Let $\mathcal{H}$ be a Hilbert space and $\xi$ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\| \leq M < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}[\|\xi\|^2]$. Let $\{\xi_i\}_{i=1}^n$ be independent random drawers of a random variable $\xi$. Then, with probability at least $1 - \delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n[\xi_i - \mathbb{E}[\xi_i]]\right\| \leq \frac{2M\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2(\xi)\log(2/\delta)}{n}}.$$

**Lemma 12** (Consistency of $\hat{g}$, $\widehat{T}$, and $\widehat{T}^*$)**.** *Suppose that Assumptions 6 and 7 holds. Let $\sigma_g^2$ and $\sigma_T^2$ be defined by*

$$\sigma_g^2 := \mathbb{E}[\|Y\varphi(A,X,Z)\|^2], \quad \sigma_T^2 := \mathbb{E}[\|\phi(A,X,W)\|^2\|\varphi(A,X,Z)\|^2].$$

*Then, each of the following statements holds true with probability at least $1-\delta$:*

$$\begin{aligned}
\|\hat{g}-g\| &\leq \frac{2c_Y\kappa^3\log(2/\delta)}{n} + \sqrt{\frac{2\sigma_g^2\log(2/\delta)}{n}} \\
\|\widehat{T}-T\| &\leq \frac{2\kappa^6\log(2/\delta)}{n} + \sqrt{\frac{2\sigma_T^2\log(2/\delta)}{n}} \\
\|\widehat{T}^*-T^*\| &\leq \frac{2\kappa^6\log(2/\delta)}{n} + \sqrt{\frac{2\sigma_T^2\log(2/\delta)}{n}}
\end{aligned}$$

*Proof.* Let $\xi_g(a,x,y,z) := y\varphi(a,x,z)$. It follows from Assumptions 6 and 7 that

$$\|\xi_g(a,x,y,z)\| \leq |y|\|\varphi(a,x,z)\| = |y|\sqrt{k(a,a)k(x,x)k(z,z)} \leq c_Y\kappa^3.$$

Hence, we have

$$\hat{g} = \frac{1}{n}\sum_{i=1}^n \xi_g(a_i,x_i,y_i,z_i), \quad g = \mathbb{E}[\xi_g(A,X,Y,Z)].$$

If $\sigma_g^2 = \mathbb{E}[\|\xi\|^2] = \mathbb{E}[\|Y\varphi(A,X,Z)\|^2]$, it follows from Lemma 11 that

$$\|\hat{g}-g\| \leq \frac{2c_Y\kappa^3\log(2/\delta)}{n} + \sqrt{\frac{2\sigma_g^2\log(2/\delta)}{n}}$$

with probability at least $1-\delta$. Next, to bound $\|\widehat{T}-T\|$, recall that we can express $\widehat{T}$ and $T$ as elements of $\mathcal{H}_{AXW}\otimes\mathcal{H}_{AXZ}$ as follows:

$$\widehat{T} = \frac{1}{n}\sum_{i=1}^n \phi(a_i,x_i,w_i)\otimes\varphi(a_i,x_i,z_i), \quad T = \int \phi(A,X,W)\otimes\varphi(A,X,Z)\,d\rho(A,X,W,Z).$$

Let $\xi_T(a,x,w,z) := \phi(a,x,w)\otimes\varphi(a,x,z) \in \mathcal{H}_{AXW}\otimes\mathcal{H}_{AXZ}$. Then, by Assumption 7,

$$\|\xi_T(a,x,w,z)\| = \|\phi(a,x,w)\|\|\varphi(a,x,z)\| \leq \sqrt{k(a,a)k(x,x)k(w,w)}\sqrt{k(a,a)k(x,x)k(z,z)} \leq \kappa^6. \tag{94}$$

As a result, we can express $\widehat{T}$ and $T$ as

$$\widehat{T} = \frac{1}{n}\sum_{i=1}^n \xi_T(a_i,x_i,w_i,z_i), \quad T = \mathbb{E}[\xi_T(A,X,W,Z)]. \tag{95}$$

Letting $\sigma_T^2 := \mathbb{E}[\|\xi_T\|^2] = \mathbb{E}[\|\phi(A,X,W)\|^2\|\varphi(A,X,Z)\|^2]$ and applying Lemma 11 yields with probability at least $1-\delta$

$$\|\widehat{T}-T\| \leq \|\widehat{T}-T\|_{\mathcal{H}_{AXW}\otimes\mathcal{H}_{AXZ}} \leq \frac{2\kappa^6\log(2/\delta)}{n} + \sqrt{\frac{2\sigma_T^2\log(2/\delta)}{n}}. \tag{96}$$

The bound on $\|\widehat{T}^*-T^*\|$ can be obtained using similar proof techniques, so we omit it for brevity. □

**Lemma 13.** $\|\widehat{T}^*\widehat{T}-T^*T\| = \mathcal{O}(1/\sqrt{n})$.

*Proof.* First, we have

$$\begin{aligned}
\|\widehat{T}^*\widehat{T}-T^*T\| &= \|\widehat{T}^*\widehat{T}-\widehat{T}^*T+\widehat{T}^*T-T^*T\| \\
&\leq \|\widehat{T}^*\widehat{T}-\widehat{T}^*T\| + \|\widehat{T}^*T-T^*T\| \\
&= \|\widehat{T}^*(\widehat{T}-T)\| + \|(\widehat{T}^*-T^*)T\| \\
&\leq \|(\widehat{T}^*-T^*)(\widehat{T}-T)\| + \|T^*(\widehat{T}-T)\| + \|(\widehat{T}^*-T^*)T\| \\
&\leq \|\widehat{T}-T\|^2 + 2\|T\|\|\widehat{T}-T\|.
\end{aligned}$$

Hence, the rate of convergence of $\|\widehat{T}^*\widehat{T} - T^*T\|$ is dominated by the rate of $\|\widehat{T} - T\|$ which, according to Lemma 12, is in the order of $\mathcal{O}(1/\sqrt{n})$. □

**Lemma 14.** $\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\| = \mathcal{O}(1/\sqrt{n})$.

*Proof.* First, we have

$$\begin{aligned}
\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\|_{\mathcal{H}_{AXW}} &= \|(\widehat{T}^*\hat{g} - T^*Th_0) + (T^*Th_0 - \widehat{T}^*\widehat{T}h_0)\|_{\mathcal{H}_{AXW}} \\
&= \|(\widehat{T}^*\hat{g} - T^*g) + (T^*Th_0 - \widehat{T}^*\widehat{T}h_0)\|_{\mathcal{H}_{AXW}} \\
&= \|(\widehat{T}^*\hat{g} - \widehat{T}^*g) + (\widehat{T}^*g - T^*g) + (T^*Th_0 - \widehat{T}^*\widehat{T}h_0)\|_{\mathcal{H}_{AXW}} \\
&\leq \underbrace{\|\widehat{T}^*\hat{g} - \widehat{T}^*g\|_{\mathcal{H}_{AXW}}}_{(A)} + \underbrace{\|\widehat{T}^*g - T^*g\|_{\mathcal{H}_{AXW}}}_{(B)} + \underbrace{\|T^*Th_0 - \widehat{T}^*\widehat{T}h_0\|_{\mathcal{H}_{AXW}}}_{(C)}.
\end{aligned}$$

Next, we will bound each term separately.

**Probabilistic bound on** $(A)$**.** Since $\widehat{T}^*$ is a Hilbert-Schmidt operator in $\mathcal{H}_{AXZ} \otimes \mathcal{H}_{AXW}$, we have by Assumption 7 that $\|\widehat{T}^*\| \leq \|\widehat{T}^*\|_{\text{HS}} \leq \kappa^3$. Consequently, $\|\widehat{T}^*\hat{g} - \widehat{T}^*g\|_{\mathcal{H}_{AXW}} = \|\widehat{T}^*(\hat{g} - g)\|_{\mathcal{H}_{AXW}} \leq \|\widehat{T}^*\|\|\hat{g} - g\|_{\mathcal{H}_{AXZ}} \leq \kappa^3\|\hat{g} - g\|_{\mathcal{H}_{AXZ}}$. By Lamma 12, we have with probability at least $1 - \delta$,

$$(A) \leq \frac{2c_Y \log(2/\delta)}{n} + \frac{1}{\kappa^3}\sqrt{\frac{2\sigma_g^2 \log(2/\delta)}{n}}. \tag{97}$$

That is, $(A) = \mathcal{O}(1/\sqrt{n})$.

**Probabilistic bound on** $(B)$**.** Using Lemma 12, we have $\|\widehat{T}^*g - T^*g\| \leq \|\widehat{T}^* - T^*\|\|g\|_{\mathcal{H}_{AXZ}} = \mathcal{O}(1/\sqrt{n})$.

**Probabilistic bound on** $(C)$**.** $\|T^*Th_0 - \widehat{T}^*\widehat{T}h_0\|_{\mathcal{H}_{AXW}} \leq \|T^*T - \widehat{T}^*\widehat{T}\|\|h_0\|_{\mathcal{H}_{AXW}} = \mathcal{O}(1/\sqrt{n})$ by Lemma 13.

Since $(A)$, $(B)$, and $(C)$ are all in the order of $\mathcal{O}(1/\sqrt{n})$, $\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\| = \mathcal{O}(1/\sqrt{n})$ as required. □

**Probabilistic bound on** $\|\widehat{\Gamma}\|$**.** Assume $\lambda \leq \|T^*T\|$ and $n \geq 2C_\epsilon\kappa\mathcal{N}(\lambda)\lambda^{-1}$. Then, with probability at least $1 - \epsilon/3$, $\|\widehat{\Gamma}\| \leq 1/\lambda$.

*Proof.* Assume

$$\|(T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda I)^{-1}\| \leq \frac{1}{2}. \tag{98}$$

Using the Neumann series of $I - (T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda)^{-1}$, we have

$$\begin{aligned}
(\widehat{T}^*\widehat{T} + \lambda I)^{-1} &= (T^*T + \lambda I)^{-1}(I - (T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda I)^{-1})^{-1} \\
&= (T^*T + \lambda I)^{-1}\sum_{k=0}^{\infty}((T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda I)^{-1})^k.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|(\widehat{T}^*\widehat{T} + \lambda I)^{-1}\| &= \|(T^*T + \lambda I)^{-1}\|\sum_{k=0}^{\infty}\|(T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda I)^{-1}\|^k \\
&\leq \|(T^*T + \lambda I)^{-1}\|\frac{1}{1 - \|(T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda I)^{-1}\|} \\
&\leq 2\|(T^*T + \lambda I)^{-1}\|.
\end{aligned}$$

where the last inequality results from (98). On the other hand, by the spectral theorem,

$$\|(T^*T + \lambda I)^{-1}\| = \sup_{l \in (l_k)_{k=0}^{\infty}} \frac{1}{l + \lambda} \leq \frac{1}{\lambda},$$

where $(l_k)_{k=0}^{\infty}$ are the eigenvalues of $T^*T$. We now prove (98). We have

$$\|(T^*T - \widehat{T}^*\widehat{T})(T^*T + \lambda I)^{-1}\| \leq \|(T^*T - \widehat{T}^*\widehat{T})\|\|(T^*T + \lambda I)^{-1}\| \leq \frac{\|(T^*T - \widehat{T}^*\widehat{T})\|}{\lambda}$$

The last term is smaller than $\mathcal{O}(1/\sqrt{n})$ with high probability. □

FINAL STEP

We have shown that $\|\widehat{T}^*\widehat{T} - T^*T\| = \mathcal{O}(1/\sqrt{n})$ and $\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\| = \mathcal{O}(1/\sqrt{n})$, and are now in a position to provide the rate of convergence of the estimation bias $\|\hat{h}_\lambda - h_0\|$.

### C.5 Proof of Theorem 3

**Theorem statement.** Suppose that $h_0 \in \mathcal{H}_\gamma$ for some $\gamma > 0$ and the conditions of Lemma 12, 13, and 14 hold. If $n^{\frac{1}{2} - \frac{1}{2}\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}$ is bounded away from zero, and $\lambda = n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}$, then

$$\|\hat{h}_\lambda - h_0\| = \mathcal{O}\left(n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}\right). \tag{99}$$

*Proof.* Suppose that $\|\widehat{T}^*\widehat{T} - T^*T\| = \mathcal{O}(1/\alpha_n)$ and $\|\widehat{T}^*\hat{g} - \widehat{T}^*\widehat{T}h_0\| = \mathcal{O}(1/\beta_n)$. Then, it follows from Proposition 13 and Carrasco et al. (2007, Proposition 4.1) that

$$\|\hat{h}_\lambda - h_0\| = \mathcal{O}\left(\frac{1}{\lambda\beta_n} + \left(\frac{1}{\lambda\alpha_n} + 1\right)\|h_\lambda - h_0\|\right). \tag{100}$$

Hence, $\lambda\beta_n$ must go to infinity as least as fast as $\|h_\lambda - h_0\|^{-1}$. That is, for $h_0 \in \mathcal{H}_\gamma$, Proposition 12 implies that

$$\lambda^2\beta_n^2 \geq \lambda^{-\min(\gamma,2)} \Rightarrow \lambda \geq \beta_n^{-\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}. \tag{101}$$

Thus, to get the fastest possible rate, we will choose $\lambda = \beta_n^{-\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}$. Consequently, the rate of convergence of $\|\hat{h}_\lambda - h_0\|$ and $\|h_\lambda - h_0\|$ will coincide if and only if $\alpha_n\beta_n^{-\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}$ is bounded away from zero. Finally, by Lemma 13 and Lemma 14, we substitute $\alpha_n = \sqrt{n}$ and $\beta_n = \sqrt{n}$ to get the stated result. □

**Proof of Proposition 3**

*Proof.* We can adapt Lemma 9 easily to see that, setting $m = n_t$ and for simplicity of notation writing $\hat{\beta} = \hat{\beta}(n_t)$,

$$\begin{aligned}|\hat{\beta}(a) - \beta(a)| &\leq \|\hat{h}_\lambda\|_{\mathcal{H}_{\mathcal{AWX}}}\|\phi(a)\|_{\mathcal{H}_\mathcal{A}}\|\hat{\mu}^{n_t} - \mu\|_{\mathcal{H}_{\mathcal{AW}}} + \|\hat{h}_\lambda - h_0\|_{\mathcal{H}_{\mathcal{AWX}}}\|\phi(a)\|_{\mathcal{H}_\mathcal{A}}\|\mu\|_{\mathcal{H}_{\mathcal{AW}}} \\ &= \mathcal{O}(\|\hat{\mu}^{n_t} - \mu\|_{\mathcal{H}_{\mathcal{AW}}}) + \mathcal{O}(\|\hat{h}_\lambda - h_0\|_{\mathcal{H}_{\mathcal{AWX}}})\end{aligned} \tag{102}$$

From Lemma 8, by setting $g$ to be the feature map on $\mathcal{A} \times \mathcal{W}$, we have

$$\|\hat{\mu}^{n_t} - \mu\| \leq \sqrt{\frac{C_k}{n_t}} + \sqrt{\frac{2C_k\log(1/\delta)}{n_t}} = \mathcal{O}(n_t^{-\frac{1}{2}}) \tag{103}$$

By Theorem 3 we have

$$\|\hat{h}_\lambda - h_0\| = \mathcal{O}\left(n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}\right). \tag{104}$$

Thus, collecting rates of both terms in (102) we get

$$|\hat{\beta}(a) - \beta(a)| = \mathcal{O}(n_t^{-\frac{1}{2}} + n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2}, \frac{1}{2}\right)}) \tag{105}$$

□

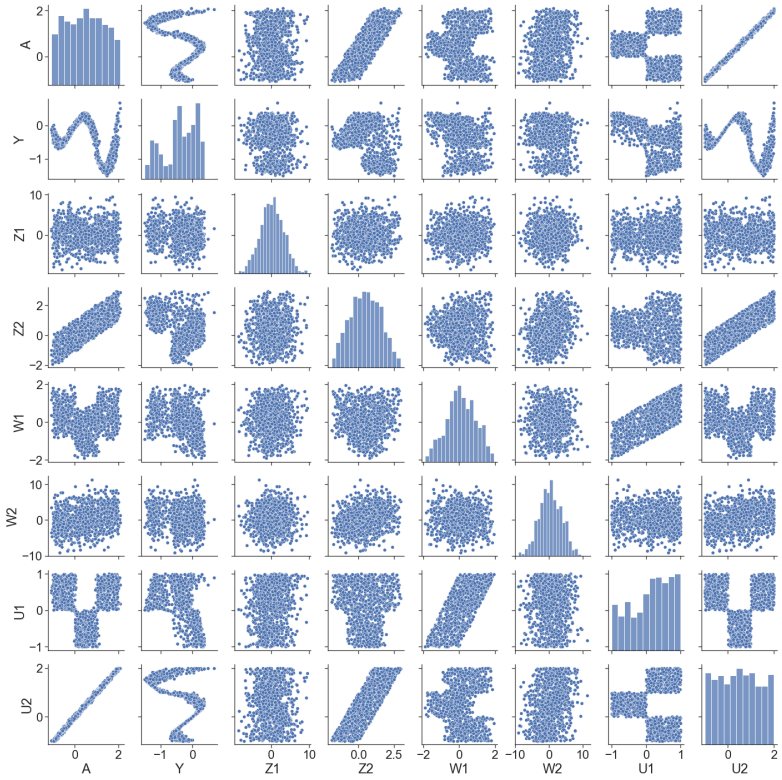


Figure 7: Synthetic generative model, sample size=1000

# D Experiments

## D.1 Data

### D.1.1 Real world data

**Disclaimer: We have applied our proposed methodologies on real world datasets to demonstrate performance of our methods. The results should only be interpreted within framework of assessing methodologies.**

For the Abortion and Criminality data (Woody et al., 2020), the treatment variable is *effective abortion rate*, the outcome variable is *murder rate*, and the covariates are *prisoner population per capita*, *state unemployment rate*, *income per capita*, *state poverty rate*, *beer consumption per capita*, *presence of concealed weapons law*, *police employment rate per capita*, and *generosity to Aid to Families with Dependent Children*. How we selected the proxy variables are described below. We mask the rest of the variables as unobserved confounders.

For the Education case study (Deaner, 2018; Fruehwirth et al., 2016), we are interested in the effect of grade retention on long-term cognitive outcome, measured in terms of a reading and maths score when the subject is aged around 11. In particular, our treatment variables are discrete, with levels at 0 (no retention), 1 (kindergarten retention) and 2 (early elementary school retention). Following (Deaner, 2018), we use as proxy variables Kindergarten test scores ($W$) and early or late elementary school test scores ($Z$). Like in the Abortion and Criminality data, we mask the rest of the variables as unobserved confounders.

To construct the *True Average Causal Effect* for real world datasets, where we do not have access to the full generative model to infer $\mathbb{E}(y|do(a))$, we have developed an empirical model to learn the *latent variable* for each dataset. Specifically, we followed the procedure below to model the latent confounder.

1. We identified the potential candidates for proxies $W$ and $Z$ by stratifying variables based on the domain knowledge and correlation with $y$ and $a$. For Criminology case study ("Legalized abortion and crime"), we followed (Woody et al., 2020) to identify proxy variables and categorize them as $W$ and $Z$. For the Educational case study ("Grade retention and Cognitive outcome"), we selected proxies as proposed by (Deaner, 2018). By this, we constructed a multi-dimensional proxy variables $W$ and $Z$ for each example.

2. We have included all other covariates as common endogenous confounders in generative model, i.e. $X$.

   In Criminology case, as proposed by (Woody et al., 2020), we added a set of exogenous common confounders to the model. In contrast with endogenous confounders, the common latent confounder ($U$) is not a parent of the exogenous confounders/covariates.

3. Assuming a generative model consistent with graph in fig. 1, we learned parameters of this generative model from data. Specifically, we assumed a graph $\mathcal{G}$ consistent with fig. 1 and learned the *Structural causal model (SCM)*, $E(V_i|pa(V_i)) = f_i(pa(V_i)), \quad \forall i \in \mathcal{G}$, for each endogenous variable. We fit a *generalized additive model* for each experiment to learn parameters of the generative model.

4. To learn the generative distribution of the unmeasured confounder, we fit a *Gaussian Mixture Model* on noise term of SCMs, learned at the previous stage. That is, we assumed the latent confounder $U$ (multidimensional confoudner unaccounted for in previous step) manifest as correlated noises of SCMs. We learned the parameters of a Gaussian Mixture model representing this latent variable.

5. We proceed to generate samples $\{(a, x, z, w, y)_i\}_{i=1}^n$ from the generative model for $\mathcal{G}$ learnt in previous steps (n=10000).

6. The True Average Causal Effect at a given $A = a$ is estimated by fixing $A$ at $a$ and averaging the $Y$ samples sampled from the fixed $A$ and the rest of its parents.

### D.2 Hyperparameters selection

For both KPV and PMMR, we employ Gaussian kernel (106) for continuous variables, as it is a continuous, bounded, and characteristic kernel and meets all assumptions required to guarantee consistency of the solution at population level.

$$k_{x_i,x_j} = exp\{-\frac{\|x_i - x_j\|^2}{2\sigma^2}\} \tag{106}$$

See (Sriperumbudur et al., 2011) for survey of properties of these kernels. For multidimensional inputs, we use the product of scalar kernels for each dimension as the kernel of the input. In both KPV and PMMR settings, we deal with two categories of hyper-parameters: (1) Kernel's length-scale ($\sigma$), and (2) regularization hyper-parameters.

#### D.2.1 Hyperparameter selection procedure (KPV).

*Kernel's length-scale*. A convenient heuristic is to set the length-scale equal to the median inter-point distances of all points in sample with size $n$. that is, $\sigma := Med(|x_i - x_j|_{\mathcal{H}}) \quad \forall i, j \in n$. We initiated the length-scale hyperparameter according to this heuristic for every input (and every dimension of multidimensional inputs). We, subsequently, chose the optimal length-scale from a narrow range around this level to allow for narrower/wider kernels to be considered.

*Regularization hyper-parameters*. For the regularization parameters, for both Stage 1 and Stage 2, we use the leave-one-out cross validation method and follow the procedure proposed in (Singh et al., 2020, Algorithm. H1) to find the optimal regularization hyper-parameter. In particular, we constructed $H_\lambda$ and $\tilde{H}_\lambda$ for Stage 1 as:

$$H_{\lambda_1} = I - \mathcal{K}_{AXZ}(\mathcal{K}_{AXZ} + m_1\lambda_1)^{-1}, \qquad \tilde{H}_{\lambda_1} = diag(H_{\lambda_1}), \quad \mathcal{K}_{AXZ} := K_{AA} \odot K_{XX} \odot K_{ZZ}$$

and implemented a grid search over $\Lambda_1$ to find $\lambda_1$ as a minimizer of the closed form of validation loss (107).

$$\hat{\lambda}_1 = \underset{\lambda_1 \in \Lambda_1}{\operatorname{argmin}} \frac{1}{m_1} |\tilde{H}_{\lambda_1}^{-1} H_{\lambda_1} K_{WW} H_{\lambda_1} \tilde{H}_{\lambda_1}^{-1}|_2, \quad \Lambda \in \mathbb{R} \tag{107}$$

For Stage 2:

$$H_{\lambda_2} = I - A\left(m_2\lambda_2 + \Sigma\right)^{-1}, \qquad \tilde{H}_{\lambda_2} = diag(H_{\lambda_2})$$

where $A := \Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} \overline{\otimes} I_{m_2 \times m_2}$ and $\Sigma$ is defined as (59). We implemented a grid search over $\Lambda_2$ to find $\lambda_2$ as a minimizer of the closed form of validation loss (108).

$$\hat{\lambda}_2 = \underset{\lambda_2 \in \Lambda_2}{\operatorname{argmin}} \frac{1}{m_2} |\tilde{H}_{\lambda_2}^{-1} H_{\lambda_2} y|_2^2, \quad \Lambda_2 \in \mathbb{R} \tag{108}$$

Note that in our setting, we assumed that the optimal hyperparameters of the first and second stages can be selected independently. In reality, however, the hyperparameter selected in first stage, has a direct effect on second stage loss and consequently, the optimal value of the hyperparameter in second stage.

#### D.2.2 HYPERPARAMETER SELECTION PROCEDURE (PMMR).

*Kernel's length-scale*. We select $\sigma_l$ and $\sigma_k$ using the median interdistance heuristic on the joint kernels $l : (A\times X\times W)^2 \to \mathbb{R}$ and $k : (A \times X \times Z)^2 \to \mathbb{R}$.

*Regularization hyper-parameters*. For the regularization parameter $\lambda$, we let $b_l^2 = (\lambda n^2)^{-1} \implies \lambda = \frac{1}{(b_l n)^2}$. For all training sizes $n$, we fixed the range of $b_l n$ to be $[2, 450]$, which translate to a range in $\lambda$ of $[4.9 \times 10^{-6}, 0.25]$, and we do grid search with a grid size of $50$.

The metric we use for hyperparameter selection is the empirical estimate of the $V - statistic$, that is, $\widehat{R}_V$. We select the hyperparameter $\lambda$ which minimizes $\hat{R}_V$ over a held-out validation set.

### D.3 Results

#### D.3.1 ABORTION & CRIMINALITY

The unobserved confounding variables ($U$) are selected as "income per capita", "police employment rate per capita", "state unemployment rate" and "state poverty rate"; the outcome inducing proxies ($W$) are selected as "prisoner population per capita", "prescence of concealed weapons law", "beer consumption per capita". We calculate their Canonical Correlation, obtaining an absolute correlation value ($|r_{CCA}|$) of $0.48$, suggesting strong correlation between $W$ and $U$.

## E A Connection between Two-stage Procedure and Maximum Moment Restrictions for the Proxy Setting

Note that $R$ and $\tilde{R}$, true loss for PMMR and KPV methods, respectively, are both positive quantities.

**Lemma 15.** *A minimizer of $\tilde{R}$ is a minimizer of $R$; and vice-versa. This minimize is unique.*

*Proof.* For any $h, h' \in L^2_{P_{AXW}}$, by developing the squares and using the law of iterated expectation, we have :

$$\begin{aligned}
\tilde{R}(h) - \tilde{R}(h') &= \mathbb{E}_{AXYZ}[(Y - \mathbb{E}[h(A,X,W) \mid A,X,Z])^2] - \mathbb{E}_{AXYZ}[(Y - \mathbb{E}[h'((A,X,W) \mid A,X,Z])^2] \\
&= 2\mathbb{E}_{AXYZ}[Y\mathbb{E}[h'(A,X,W) - h(A,X,W)|A,X,Z]] + \mathbb{E}_{AXZ}[\mathbb{E}[h(A,X,W)|A,X,Z]^2] \\
&\quad - \mathbb{E}_{AXZ}[\mathbb{E}[h'(A,X,W)|A,X,Z]^2] \\
&= 2\mathbb{E}_{AXZ}[\mathbb{E}[Y|A,X,Z]\mathbb{E}[h'(A,X,W) - h(A,X,W)|A,X,Z]] \\
&\quad + \mathbb{E}_{AXZ}[\mathbb{E}[h(A,X,W)|A,X,Z]^2] - \mathbb{E}_{AXZ}[\mathbb{E}[h'(A,X,W)|A,X,Z]^2] \\
&= R(h) - R(h').
\end{aligned}$$

Assuming $\exists h, h' \in L^2_{P_{A,X,W}}$ such that $\mathbb{E}[Y|A,X,Z] = \mathbb{E}[h(A,X,W)|A,X,Z]$, according to the preceding computations we have:

$$\begin{aligned}
\tilde{R}(h) - \tilde{R}(h') &= R(h) - R(h') \\
&= 2\mathbb{E}_{AXZ}[\mathbb{E}[Y|A,X,Z]\mathbb{E}[h'(A,X,W) - h(A,X,W)|A,X,Z]] \\
&\quad + \mathbb{E}_{AXZ}[\mathbb{E}[h(A,X,W)|A,X,Z]^2] - \mathbb{E}_{AXZ}[\mathbb{E}[h'(A,X,W)|A,X,Z]^2] \\
&= \mathbb{E}_{AXZ}[\mathbb{E}[h'(A,X,W)|A,X,Z]^2] - 2\mathbb{E}_{AXZ}[\mathbb{E}[h(A,X,W)|A,X,Z]\mathbb{E}[h'(A,X,W)|A,X,Z]] \\
&\quad + \mathbb{E}_{AXZ}[\mathbb{E}[h(A,X,W)|A,X,Z]^2] \\
&= \mathbb{E}_{AXZ}[(\mathbb{E}[h'(A,X,W)|A,X,Z] - \mathbb{E}[h(A,X,W)|A,X,Z])^2].
\end{aligned}$$

Taking $h' = h$ in the equation above shows that $h$ is a minimizer of $R$ and $\tilde{R}$ Hence, a unique minimizer of $R$ is a minimizer of $\tilde{R}$; and vice-versa. □

2. By Lemma 1, $R(h) = 0$ if and only if $h$ satisfies the conditional moment restriction (CMR): $\mathbb{E}[Y - h(A,W,X) \mid A,Z,X] = 0$, $\mathbb{P}(A,Z,X)$-almost surely. We now show $R_k(h) = 0$ if and only if $R(h) = 0$. Firstly, by the law of iterated expectations,

$$\begin{aligned}
\mathbb{E}[(Y - h(A,W,X))k((A,Z,X),\cdot)] &= \mathbb{E}_{A,X,Z}[\mathbb{E}[(Y - h(A,W,X))k((A,Z,X),\cdot)|A,X,Z]] \\
&= \mathbb{E}_{A,X,Z}[\mathbb{E}[(Y - h(A,W,X))|A,X,Z]k((A,Z,X),\cdot)].
\end{aligned}$$

By Lemma 2, $R_k(h) = \|\mathbb{E}[(Y - h(A,W,X))k((A,Z,X),\cdot)]\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{Z}\mathcal{X}}}$. Hence, if $h$ satisfies the CMR condition, then $R_k(h) = 0$. We now assume that $R_k(h) = 0$. We can write $R_k(h)$ as:

$$\iint g(a,x,z)k((a,x,z),(a',x',z')g(a',x',z')d(a,x,z)d(a,x,z) = 0,$$

where we define $g(a,x,z) = \mathbb{E}_{AXWY}[Y - h(A,X,W)|a,x,z]d\rho(a,x,z)$. Since $k$ is ISPD by Assumption 11, this implies the CMR: $\mathbb{E}[Y - h(A,W,X) \mid A,Z,X] = 0$, $\mathbb{P}(A,Z,X)$-almost surely.

3. In KPV, the method is decomposed in two stages.

*First stage.* Under the assumption that $\mathbb{E}[f(w)|A,X,Z=\cdot]$ is in $\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}}$ for any $f \in \mathcal{H}_W$, the conditional mean embedding $\mu$ can be written $\mu_{W|a,x,z} = C_{W|A,X,Z}\phi(a,x,z)$ for any $(a,x,z) \in \mathcal{A}\times\mathcal{X}\times\mathcal{Z}$, where $C_{W|A,X,Z} : \mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}} \to \mathcal{H}_{\mathcal{W}}$ is the conditional mean embedding operator is well-defined (Song et al., 2009). Let $\mathcal{H}_\Gamma$ the vector-valued RKHS of operators from $\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}}$ to $\mathcal{H}_{\mathcal{W}}$. A crucial result is that the tensor product $\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}} \otimes \mathcal{H}_{\mathcal{W}}$ is isomorphic to $\mathcal{L}^2(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$ the space of Hilbert-Schmidt operators from $\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}}$ to $\mathcal{H}_{\mathcal{W}}$. Hence, by choosing the vector-valued kernel $\Gamma$ with feature map : $(w,a,x,z) \mapsto [\phi(a)\otimes\phi(x)\otimes\phi(z)\otimes\phi(w)] = \phi(a)\otimes\phi(x)\otimes\phi(z)\langle\phi(w),\cdot\rangle_{\mathcal{H}_{\mathcal{W}}}$, we have $\mathcal{H}_\Gamma = \mathcal{L}^2(\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$ and they share the same norm. We denote by $L^2(\mathcal{A}\times\mathcal{X}\times\mathcal{Z}, \rho_{\mathcal{A}\mathcal{X}\mathcal{Z}})$ the space of square integrable functions from $\mathcal{A}\times\mathcal{X}\times\mathcal{Z}$ to $\mathcal{W}$ with respect to measure $\rho_{\mathcal{A}\mathcal{X}\mathcal{Z}}$, where $\rho_{\mathcal{A}\mathcal{X}\mathcal{Z}}$ is the restriction of $\rho$ to $\mathcal{A}\times\mathcal{X}\times\mathcal{Z}$. Assuming $C_{W|A,X,Z} \in \mathcal{H}_\Gamma$, it is the solution to the following risk minimization:

$$C_{W|A,X,Z} = \operatorname*{argmin}_{c\in\mathcal{H}_\Gamma} E(C) \quad \text{where} \quad E(C) = \mathbb{E}_{AXZW}\left[\|\phi(W) - C\phi(a,x,z)\|^2_{\mathcal{H}_{\mathcal{W}}}\right] \tag{109}$$

*Second stage.* Under the assumptions of a characteristic kernel and that $h_0 \in \mathcal{H}_{AW}$, $\mathbb{E}[h(A,X,W)|A,X,Z] = \eta_{AXW}[\phi(a,x)\otimes\mu_{W|a,x,z}]$. The operator $\eta_{AXW}$ minimizes

$$\eta_{AXW} = \operatorname*{argmin}_{\eta\in\mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}} \tilde{R}(\eta) \quad \text{where} \quad \tilde{R}(\eta) = \mathbb{E}_{AXYZ}\left[(Y - \eta_{AXW}[\phi(a,x)\otimes\mu_{W|a,x,z}])^2\right],$$

where $\mu_{W|a,x,z} = C_{W|A,X,Z}\phi(a,x,z)$ and $C_{W|A,X,Z}$ is the solution of (109). Hence, as long as the problem is well-posed, i.e $C_{W|A,X,Z} \in \mathcal{H}_\Gamma$ and $h \in \mathcal{H}_{\mathcal{A}\mathcal{X}\mathcal{W}}$, the KPV approach recovers $\mathbb{E}[h(A,X,W)|\cdot]$, with $\mathbb{E}[h(A,X,W)|A,X,Z] = \eta_{AXW}[\phi(a,x)\otimes\mu_{W|A,X,Z}] = \eta_{AXW}[\phi(a,x)\otimes C_{W|A,X,Z}\phi(A,X,Z)]$.