A. Background Material

In this section, we give a more detailed review of the background on reproducing kernel Hilbert space embeddings and U-statistics. Interested readers can refer to Berlinet & Thomas-Agnan (2004); Muandet et al. (2017) for the former, and Serfling (1980, Chapter 5) for the latter.

A.1. Reproducing Kernel Hilbert Space Embeddings

Let $\mathcal{H}$ be a vector space of real-valued functions on $\mathcal{Y}$, endowed with the structure of a Hilbert space via an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Let $\|\cdot\|_{\mathcal{H}}$ be the associated norm, i.e. $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{\frac{1}{2}}$ for $f \in \mathcal{H}$.

Definition A.1 (Berlinet & Thomas-Agnan (2004, p.7, Definition 1))**.** A function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a *reproducing kernel* of the Hilbert space $\mathcal{H}$ if and only if

(i) for all $y \in \mathcal{Y}$, $l(y, \cdot) \in \mathcal{H}$;

(ii) for all $y \in \mathcal{Y}$ and for all $f \in \mathcal{H}$, $\langle f, l(y, \cdot) \rangle_{\mathcal{H}} = f(y)$ (the *reproducing property*).

A Hilbert space of functions $\mathcal{Y} \to \mathbb{R}$ which possesses a reproducing kernel is called the *reproducing kernel Hilbert space* (RKHS).

For any $y \in \mathcal{Y}$, denote by $e_y : \mathcal{H} \to \mathbb{R}$ the evaluation functional at y, i.e. $e_y(f) = f(y)$ for $f \in \mathcal{H}$. Riesz representation theorem can be used to prove the following lemma.

Lemma A.2 (Berlinet & Thomas-Agnan (2004, p.9, Theorem 1))**.** *A Hilbert space of functions $\mathcal{Y} \to \mathbb{R}$ has a reproducing kernel if and only if all evaluation functionals $e_y, y \in \mathcal{Y}$ are continuous on $\mathcal{H}$.*

Next, we characterise reproducing kernels.

Definition A.3 (Berlinet & Thomas-Agnan (2004, p.10, Definition 2))**.** A function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is called a *positive definite function* if, for all $n \geq 1$, any $a_1, ..., a_n \in \mathbb{R}$ and any $y_1, ..., y_n \in \mathcal{Y}$,

$$\sum_{i,j=1}^{n} a_i a_j l(y_i, y_j) \geq 0.$$

A reproducing kernel is a positive definite function, since, by the reproducing property,

$$\sum_{i,j=1}^{n} a_i a_j l(y_i, y_j) = \left\| \sum_{i=1}^{n} a_i l(y_i, \cdot) \right\|_{\mathcal{H}}^2 \geq 0$$

(see Berlinet & Thomas-Agnan (2004, p.13, Lemma 2)). The Moore-Aronszajn Theorem (Aronszajn, 1950) shows that the set of positive definite functions and the set of reproducing kernels on $\mathcal{Y} \times \mathcal{Y}$ are identical.

Theorem A.4 (Berlinet & Thomas-Agnan (2004, p.19, Theorem 3))**.** *Let l be a positive definite function on $\mathcal{Y} \times \mathcal{Y}$. Then there exists a unique Hilbert space of functions $\mathcal{Y} \to \mathbb{R}$ with l as its reproducing kernel. The subspace $\tilde{\mathcal{H}}$ of $\mathcal{H}$ spanned by $\{l(y, \cdot) : y \in \mathcal{Y}\}$ is dense in $\mathcal{H}$, and $\mathcal{H}$ is the set of functions $\mathcal{Y} \to \mathbb{R}$ which are pointwise limits of Cauchy sequences in $\tilde{\mathcal{H}}$ with the inner product*

$$\langle f, g \rangle_{\tilde{\mathcal{H}}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j l(y_i, y_j)$$

where $f = \sum_{i=1}^{n} \alpha_i l(y_i, \cdot)$ and $g = \sum_{j=1}^{m} \beta_j l(y_j, \cdot)$.

Examples of commonly used kernels in Euclidean spaces include the linear kernel $l(y, y') = y \cdot y'$, the monomial kernel $l(y, y') = (y \cdot y')^p$, the polynomial kernel $l(y, y') = (y \cdot y' + 1)^p$, the Gaussian kernel $l(y, y') = e^{-\frac{1}{\sigma^2}\|y - y'\|_2^2}$ and the Laplacian kernel $l(y, y') = e^{-\frac{1}{\sigma^2}\|y - y'\|_1}$.

Kernel methods in machine learning turns linear methods into non-linear ones using the so-called "kernel trick", whereby individual datapoints $y \in \mathcal{Y}$ are "embedded" into an RKHS $\mathcal{H}$ with reproducing kernel l via the mapping $y \mapsto l(y, \cdot)$. The

RKHS is high- (and often infinite-)dimensional, and performing a linear method (e.g. linear regression, support vector machine, principal component analysis, etc.) in $\mathcal{H}$ with datapoints $l(y_i, \cdot), i = 1, ..., n$, instead of the original space $\mathcal{Y}$ with datapoints $y_i, i = 1, ..., n$, results in a nonlinear method in the original space. Please see Scholkopf & Smola (2001) for more details.

Recently, this idea of RKHS embeddings has been extended to embed entire (conditional) distributions, rather than individual datapoints, via the expectation. Suppose Y is a random variable taking values in $\mathcal{Y}$, with distribution P_Y. Assuming the integrability condition $\int_{\mathcal{Y}} \sqrt{l(y,y)} dP_Y(y) < \infty$, we define the *kernel mean embedding* $\mu_{P_Y} \in \mathcal{H}$ of the measure P_Y, or the random variable Y, as

$$\mu_{P_Y}(\cdot) = \mathbb{E}\left[l(Y, \cdot)\right] = \int_{\mathcal{Y}} l(y, \cdot) dP_Y(y) = \int_{\Omega} l(Y(\omega), \cdot) dP(\omega).$$

Note that the integrand $l(Y, \cdot)$ is an element in a Hilbert space (and therefore a Banach space), so the integral is not the usual Lebesgue integral on $\mathbb{R}$. There are a number of ways in which one can define integration on a Banach space (Schwabik & Ye, 2005). Among those, the Bochner integral (Dinculeanu, 2000, p.15, Definition 35) is the simplest and most intuitive one, and suffices for our purposes. Riesz representation theorem is again used to prove the following mean embedding version of the reproducing property.

Lemma A.5 (Smola et al. (2007))**.** *For each $f \in \mathcal{Y}$,*

$$\mathbb{E}\left[f(Y)\right] = \int_{\mathcal{Y}} f(y) dP_Y(y) = \left\langle f, \mu_{P_Y} \right\rangle_{\mathcal{H}}.$$

Using the kernel mean embedding, we can define a distance function, called the *maximum mean discrepancy* (Gretton et al., 2012), between two random variables Y and Y' on $\mathcal{Y}$, or equivalently, two probability measures P_Y and $P_{Y'}$, as

$$\text{MMD}(Y, Y') = \left\| \mu_{P_Y} - \mu_{P_{Y'}} \right\|_{\mathcal{H}}.$$

The name maximum mean discrepancy comes from the following lemma.

Lemma A.6 (Gretton et al. (2012, Lemma 4))**.** *We have*

$$\text{MMD}(Y, Y') = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \left\{ \mathbb{E}\left[f(Y)\right] - \mathbb{E}\left[f(Y')\right] \right\}.$$

In this alternative definition of the MMD, the function in the unit ball of $\mathcal{H}$ that maximises $\mathbb{E}[f(Y)] - \mathbb{E}[f(Y')]$ is called the *witness function* (Gretton et al., 2012, Section 2.3). It can easily be seen that the witness function is in fact

$$\frac{\mu_{P_Y} - \mu_{P_{Y'}}}{\left\| \mu_{P_Y} - \mu_{P_{Y'}} \right\|_{\mathcal{H}}}.$$

Lloyd & Ghahramani (2015) uses the unnormalised witness function $\mu_{P_Y} - \mu_{P_{Y'}}$ for model criticism.

The MMD is not a proper metric, since Y and Y' may be distinct and still give $\text{MMD}(Y, Y') = 0$, depending on the kernel l that is used. The notion of *characteristic kernels* is therefore essential, since it tells us whether the associated RKHS is rich enough to enable us to distinguish distinct distributions based on their embeddings.

Definition A.7 (Fukumizu et al. (2008, Section 2.2))**.** Denote by $\mathcal{P}$ the set of all probability measures on $\mathcal{Y}$. A positive definite kernel l is *characteristic* if the kernel mean embedding map $\mathcal{P} \to \mathcal{H} : P_Y \mapsto \mu_{P_Y}$ is injective.

For example, of the aforementioned kernels, the Gaussian and Laplacian kernels are characteristic, whereas the linear, monomial and polynomial kernels are not. The MMD associated with a characteristic kernel is then a proper metric between probability measures on $\mathcal{Y}$. See Sriperumbudur et al. (2010; 2011); Simon-Gabriel & Schölkopf (2018) for various characterisations of characteristic kernels.

Now we discuss conditional embedding of distributions into RKHSs. Suppose X is a random variable on a space $\mathcal{X}$.

Definition A.8 (Park & Muandet (2020a, Definition 3.1))**.** The *conditional mean embedding* of the random variable Y, or equivalently, the distribution P_Y, is the Bochner conditional expectation (as defined in Dinculeanu (2000, p.45, Definition 38))

$$\mu_{P_{Y|X}} = \mathbb{E}\left[l(Y, \cdot) \mid X\right].$$

Notice that this is a straightforward extension of the kernel mean embedding $\mu_{P_Y} = \mathbb{E}[l(Y, \cdot)]$ to the conditional case.

A.2. U-Statistics

Suppose $Y_1, Y_2, ..., Y_r$ are independent copies of the random variable Y, i.e. they are independent and all have distribution P_Y. Let $h : \mathcal{Y}^r \to \mathbb{R}$ be a symmetric function (called a *kernel* in the U-statistics literature; confusion must be avoided with the reproducing kernel used throughout this paper), i.e. for any permutation π of $\{1, ..., r\}$, we have $h(y_1, ..., y_r) = h(y_{\pi(1)}, ..., y_{\pi(r)})$. Suppose we would like to estimate a function of the form

$$\theta(P_Y) = \mathbb{E}\left[h\left(Y_1, ..., Y_r\right)\right] = \int_{\mathcal{Y}} ... \int_{\mathcal{Y}} h\left(y_1, ..., y_r\right) dP_Y(y_1)...dP_Y(y_r).$$

The corresponding *U-statistic* for an unbiased estimation of $\theta(P_Y)$ based on a sample $Y_1, ..., Y_n$ of size $n \geq r$ is given by

$$\hat{\theta}(P_Y) = \frac{1}{\binom{n}{r}} \sum h\left(Y_{i_1}, ..., Y_{i_r}\right),$$

where $\binom{n}{r}$ is the binomial coefficient and the summation is over the $\binom{n}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n\}$. Clearly, since the expectation of each summand yields $\theta(P_Y)$, we have $\mathbb{E}[\hat{\theta}(P_Y)] = \theta(P_Y)$, so U-statistics are unbiased estimators.

Some examples of h and the corresponding estimator include the sample mean $h(y) = y$, the sample variance $h(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$, the sample cumulative distribution up to y^* $h(y) = \mathbf{1}(y \leq y^*)$, the k^{th} sample raw moment $h(y) = y^k$ and Gini's mean difference $h(y_1, y_2) = |y_1 - y_2|$.

To the best of our knowledge, Stute (1991) was the first to consider a conditional counterpart of U-statistics. Let $X_1, ..., X_r$ be independent copies of the random variable X. We are now interested in the estimation of the following quantity:

$$\theta\left(P_{Y|X}\right) = \mathbb{E}\left[h\left(Y_1, ..., Y_r\right) \mid X_1, ..., X_r\right].$$

By Çınlar (2011, p.146, Theorem 1.17), $\theta(P_{Y|X})$ can be considered as a function $\mathcal{X}^r \to \mathbb{R}$, such that for each r-tuple $\{x_1, ..., x_r\}$, we have

$$\theta\left(P_{Y|X}\right)(x_1, ..., x_r) = \mathbb{E}\left[h\left(Y_1, ..., Y_r\right) \mid X_1 = x_1, ..., X_r = x_r\right].$$

The simplest case is when $r = 1$ and $h(y) = y$. In this case, the estimand reduces to $f(X) = \mathbb{E}[Y|X]$, which is the usual regression problem for which a plethora of methods exist. Suppose we have a sample $\{(X_i, Y_i)\}_{i=1}^n$. One such regression method is the Nadaraya-Watson kernel smoother:

$$\hat{f}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{a}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{a}\right)},$$

where K is the so-called "smoothing kernel" and a is the bandwidth. This was extended by Stute (1991) to $r \geq 1$ and more general h:

$$\hat{\theta}\left(P_{Y|X}\right)(x_1, ..., x_r) = \frac{\sum h\left(Y_{i_1}, ..., Y_{i_r}\right) \prod_{j=1}^r K\left(\frac{x_j - X_{i_j}}{a}\right)}{\sum \prod_{j=1}^r K\left(\frac{x_j - X_{i_j}}{a}\right)},$$

where the sums are over the $\binom{n}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n\}$ as before. Derumigny (2019) considers a parametric model of the form

$$\Lambda\left(\theta\left(P_{Y|X}\right)(x_1, ..., x_r)\right) = \boldsymbol{\psi}\left(x_1, ..., x_r\right)^T \beta^*,$$

where Λ is a strictly increasing and continuously differentiable "link function" such that the range of $\Lambda \circ \theta$ is exactly $\mathbb{R}$, $\beta^* \in \mathbb{R}^s$ is the true parameter and $\boldsymbol{\psi}(\cdot) = \left(\psi_1(\cdot), ..., \psi_s(\cdot)\right)^T \in \mathbb{R}^s$ is some basis, such as polynomials, exponentials, indicator functions etc. However, the estimation of β^* still makes use of the Nadaraya-Watson kernel smoothers considered above.

Of course, Nadaraya-Watson kernel smoothers are far from being the only method of regression that can be extended to estimate conditional U-statistics, and in the main body of the paper (Section 5.2), we consider extending kernel ridge regression for this purpose.

B. More Details on IHDP Dataset

In this section, we give more details on the data generating process of the semi-synthetic IHDP (Infant Health and Development Program) dataset that was first used in the treatment effect literature by Hill (2011).

The data consists of 25 covariates: birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status, whether or not the mother smoked during pregnancy, whether or not the mother drank alcohol during pregnancy, whether or not the mother took drugs during pregnancy, the mother's age, marital status, education attainment, whether or not the mother worked during pregnancy, whether she received prenatal care, and 7 dummy variables for the 8 sites in which the family resided at the start of the intervention.

These covariates are originally taken from a randomised experiment, and included information about the ethnicity of the mothers. Hill (2011) removed all children with nonwhite mothers from the treatment group, which is clearly a non-random (biased) portion of the data, thereby imitating an observational study. This leaves 608 children in the control group and 139 in the treatment group. The overlap condition is now only satisfied for the treatment group.

In creating the parallel linear response surfaces, which are used in all three of the settings "SN", "LN" and "HN", we let $\mathbb{E}[Y_0|X] = \beta X$ and $\mathbb{E}[Y_1|X] = \beta X + 4$, where the 25-dimensional coefficient vector β is generated in the same way as in Alaa & Schaar (2018): for the 6 continuous variables (birth weight, head circumference, weeks born preterm, birth order, neonatal health index, mother's age), the corresponding coefficients is sampled from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $\{0.5, 0.125, 0.125, 0.125, 0.125\}$ respectively, whereas for the other 19 binary variables, the corresponding coefficients are sampled from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $\{0.6, 0.1, 0.1, 0.1, 0.1\}$ respectively.

Finally, we generate realisations of the potential outcomes by adding noise to the mean response surfaces. We let $Y_0 = \beta X + \epsilon(X)$ and $Y_1 = \beta X + 4 + \epsilon(X)$, where $\epsilon(X) = \epsilon_{\text{SN}}$ in setting "SN", $\epsilon(X) = \epsilon_{\text{LN}}$ in setting "LN" and $\epsilon(X) = X_6\epsilon_{\text{SN}} + (1 - X_6)\epsilon_{\text{LN}}$ in setting "HN", with $\epsilon_{\text{SN}} \sim \mathcal{N}(0, 1^2)$ and $\epsilon_{\text{LN}} \sim \mathcal{N}(0, 20^2)$. The covariate X_6 corresponds to the sex of the child, and was chosen because there are roughly the same number of each sex in both the control and the treatment groups.

C. Proofs

Lemma 4.1. *For each $x \in \mathcal{X}$, we have*

$$\hat{U}^2_{\text{MMD}}(x) = \boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\boldsymbol{k}_0(x) - 2\boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\boldsymbol{k}_1(x) + \boldsymbol{k}_1^T(x)\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\boldsymbol{k}_1(x),$$

where $[\mathbf{L}_0]_{1\le i,j\le n_0} = l(y_i^0, y_j^0)$, $[\mathbf{L}]_{1\le i\le n_0, 1\le j\le n_1} = l(y_i^0, y_j^1)$ *and* $[\mathbf{L}_1]_{1\le i,j\le n_1} = l(y_i^1, y_j^1)$.

Proof. We use the reproducing property of $\mathcal{H}$ and (2) to see that, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\hat{U}^2_{\text{MMD}}(x) &= \left\|\hat{\mu}_{Y_1|X=x} - \hat{\mu}_{Y_0|X=x}\right\|^2_{\mathcal{H}} \\
&= \left\|\boldsymbol{k}_0^T(x)\mathbf{W}_0\boldsymbol{l}_0 - \boldsymbol{k}_1^T(x)\mathbf{W}_1\boldsymbol{l}_1\right\|^2_{\mathcal{H}} \\
&= \left\langle \sum_{i,j=1}^{n_0} k_0(x, x_i^0)\mathbf{W}_{0,ij}l(y_j^0, \cdot), \sum_{p,q=1}^{n_0} k_0(x, x_p^0)\mathbf{W}_{0,pq}l(y_q^0, \cdot)\right\rangle_{\mathcal{H}} \\
&\quad - 2\left\langle \sum_{i,j=1}^{n_0} k_0(x, x_i^0)\mathbf{W}_{0,ij}l(y_j^0, \cdot), \sum_{p,q=1}^{n_1} k_1(x, x_p^1)\mathbf{W}_{1,pq}l(y_q^1, \cdot)\right\rangle_{\mathcal{H}} \\
&\quad + \left\langle \sum_{i,j=1}^{n_1} k_1(x, x_i^1)\mathbf{W}_{1,ij}l(y_j^1, \cdot), \sum_{p,q=1}^{n_1} k_1(x, x_p^1)\mathbf{W}_{1,pq}l(y_q^1, \cdot)\right\rangle_{\mathcal{H}} \\
&= \sum_{i,j,p,q=1}^{n_0} k_0(x, x_i^0)\mathbf{W}_{0,ij}l(y_j^0, y_q^0)\mathbf{W}_{0,qp}^T k_0(x_p^0, x) \\
&\quad - 2\sum_{i,j=1}^{n_0}\sum_{p,q=1}^{n_1} k_0(x, x_i^0)\mathbf{W}_{0,ij}l(y_j^0, y_q^1)\mathbf{W}_{1,qp}^T k_1(x_p^1, x)
\end{aligned}$$

$$+ \sum_{i,j,p,q=1}^{n_1} k_1(x, x_i^1)\mathbf{W}_{1,ij} l(y_j^1, y_q^1)\mathbf{W}_{1,qp}^T k_1(x_p^1, x)$$
$$= \boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\boldsymbol{k}_0(x) - 2\boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\boldsymbol{k}_1(x) + \boldsymbol{k}_1^T(x)\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\boldsymbol{k}_1(x).$$

□

Theorem 4.2. *Suppose that k_0, k_1 and l are bounded, that Γ_0 and Γ_1 are universal, and that $\lambda_{n_0}^0$ and $\lambda_{n_1}^1$ decay at slower rates than $\mathcal{O}(n_0^{-1/2})$ and $\mathcal{O}(n_1^{-1/2})$ respectively. Then as $n_0, n_1 \to \infty$,*

$$\psi_{\text{MMD}}\left(\hat{U}_{\text{MMD}}\right) = \mathbb{E}\left[\left(\hat{U}_{\text{MMD}}(X) - U_{\text{MMD}}(X)\right)^2\right] \xrightarrow{P} 0.$$

Proof. The simple inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ holds in any Hilbert space. Using this, we see that

$$\begin{aligned}
\psi_{\text{MMD}}\left(\hat{U}_{\text{MMD}}\right) &= \mathbb{E}\left[\left(\hat{U}_{\text{MMD}}(X) - U_{\text{MMD}}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}} - \left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}\right)^2\right] \\
&\leq \mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X} - \hat{\mu}_{Y_0|X} + \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] && \text{by the reverse triangle inequality} \\
&\leq 2\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2 + \left\|\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] && \text{by the above inequality.}
\end{aligned}$$

Hence, it suffices to know that

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0 \qquad \text{and} \qquad \mathbb{E}\left[\left\|\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0.$$

But this follows immediately from Park & Muandet (2020b), so the proof is complete. □

Lemma 4.3. *If l is a characteristic kernel, $P_{Y_0|X} \equiv P_{Y_1|X}$ if and only if $t = 0$.*

Proof. We can assume without loss of generality that $P_{Y_0|X}$ and $P_{Y_1|X}$ are obtained from a regular version of $P(\cdot \mid X)$. Then by (Park & Muandet, 2020a, Theorem 2.9), there exist $C_0, C_1 \in \mathcal{F}$ with $P(C_0) = P(C_1) = 1$ such that for all $\omega \in C_0$, $\mu_{Y_0|X}(\omega) = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X}(\omega)(y)$ and for all $\omega' \in C_1$, $\mu_{Y_1|X}(\omega') = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X}(\omega')(y)$.

Suppose for contradiction that there exists some measurable $A \subseteq \mathcal{X}$ with $P_X(A) > 0$ such that for all $x \in A$, $\mu_{Y_0|X=x} \neq \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y)$. Then $P(X^{-1}(A)) = P_X(A) > 0$, and hence $P(X^{-1}(A) \cap C_0) > 0$. For all $\omega \in X^{-1}(A) \cap C_0$, we have $X(\omega) \in A$, and hence

$$\mu_{Y_0|X}(\omega) \neq \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=X(\omega)}(y) = \int_{\mathcal{Y}} l(y,\cdot)P_{Y_0|X}(\omega)(dy) = \mu_{Y_0|X}(\omega).$$

This is a contradiction, hence there does not exist a measurable $A \subseteq \mathcal{X}$ with $P_X(A) > 0$ such that for all $x \in A$, $\mu_{Y_0|X=x} \neq \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y)$. Therefore, there must exist some measurable $A_0 \subseteq \mathcal{X}$ with $P_X(A_0) = 1$ such that for all $x \in A_0$, $\mu_{Y_0|X=x} = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y)$. Similarly, there must exist some measurable $A_1 \subseteq \mathcal{X}$ with $P_X(A_1) = 1$ such that for all $x \in A_1$, $\mu_{Y_1|X=x} = \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X=x}(y)$.

($\Longrightarrow$) Suppose that $P_{Y_0|X} \equiv P_{Y_1|X}$. This means that there exists a measurable $A \subseteq \mathcal{X}$ with $P_X(A) = 1$ such that for all $x \in A$, the measures $P_{Y_0|X=x}(\cdot)$ and $P_{Y_1|X=x}(\cdot)$ are the same. Then for all $x \in A \cap A_0 \cap A_1$,

$$\begin{aligned}
\mu_{Y_0|X=x} &= \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_0|X=x}(y) && \text{since } x \in A_0 \\
&= \int_{\mathcal{Y}} l(y,\cdot)dP_{Y_1|X=x}(y) && \text{since } x \in A
\end{aligned}$$

$$= \mu_{Y_1|X=x} \qquad \text{since } x \in A_1.$$

Now, we have $P_X(A) = P_X(A_0) = P_X(A_1) = 1$, so $P_X(A \cap A_0 \cap A_1) = 1$. Since $\mu_{Y_0|X=x} = \mu_{Y_1|X=x}$ for all $x \in A \cap A_0 \cap A_1$, we have $\mu_{Y_0|X=\cdot} = \mu_{Y_1|X=\cdot}$ P_X-almost everywhere. Hence,

$$t = \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] = 0$$

($\Longleftarrow$) Now suppose that $t = 0$, i.e. $\mu_{Y_0|X=\cdot} = \mu_{Y_1|X=\cdot}$ P_X-almost everywhere, say on a measurable set $A \subseteq \mathcal{X}$ with $P_X(A) = 1$. Suppose $x \in A \cap A_0 \cap A_1$. Then

$$\begin{aligned}
\int_{\mathcal{Y}} l(y,\cdot) dP_{Y_0|X=x}(y) &= \mu_{Y_0|X=x} && \text{since } x \in A_0 \\
&= \mu_{Y_1|X=x} && \text{since } x \in A \\
&= \int_{\mathcal{Y}} l(y,\cdot) dP_{Y_1|X=x}(y) && \text{since } x \in A_1.
\end{aligned}$$

Since $k_{\mathcal{Y}}$ is characteristic, this means that $P_{Y_0|X=x}$ and $P_{Y_1|X=x}$ are the same measure. As before, we have $P_X(A \cap A_0 \cap A_1) = 1$, hence $P_{Y_0|X} \equiv P_{Y_1|X}$.

□

Lemma 4.4. *We have*

$$\hat{t} = \frac{1}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_0 \mathbf{W}_0 \mathbf{L}_0 \mathbf{W}_0^T \tilde{\mathbf{K}}_0^T\right) - \frac{2}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_0 \mathbf{W}_0 \mathbf{L} \mathbf{W}_1^T \tilde{\mathbf{K}}_1^T\right) + \frac{1}{n}\mathrm{Tr}\left(\tilde{\mathbf{K}}_1 \mathbf{W}_1 \mathbf{L}_1 \mathbf{W}_1^T \tilde{\mathbf{K}}_1^T\right),$$

where $\mathbf{L}_0, \mathbf{L}_1$ *and* $\mathbf{L}$ *are as defined in Lemma 4.1 and* $[\tilde{\mathbf{K}}_0]_{1\le i\le n, 1\le j\le n_0} = k_0(x_i, x_j^0)$ *and* $[\tilde{\mathbf{K}}_1]_{1\le i\le n, 1\le j\le n_1} = k_1(x_i, x_j^1)$.

Proof. See that, using the reproducing property in $\mathcal{H}$ again,

$$\begin{aligned}
\hat{t} &= \frac{1}{n}\sum_{i=1}^n \left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left\{\left\|\hat{\mu}_{Y_1|X=x_i}\right\|_{\mathcal{H}}^2 - 2\left\langle \hat{\mu}_{Y_1|X=x_i}, \hat{\mu}_{Y_0|X=x_i}\right\rangle_{\mathcal{H}} + \left\|\hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^n \left\{\left\|\boldsymbol{k}_0^T(x_i)\mathbf{W}_0\boldsymbol{l}_0\right\|_{\mathcal{H}}^2 - 2\left\langle \boldsymbol{k}_0^T(x_i)\mathbf{W}_0\boldsymbol{l}_0, \boldsymbol{k}_1^T(x_i)\mathbf{W}_1\boldsymbol{l}_1\right\rangle_{\mathcal{H}} + \left\|\boldsymbol{k}_1^T(x_i)\mathbf{W}_1\boldsymbol{l}_1\right\|_{\mathcal{H}}^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle \sum_{p,q=1}^{n_0} k_0(x_p^0, x_i)\mathbf{W}_{0,pq} l(y_q^0,\cdot), \sum_{r,s=1}^{n_0} k_0(x_r^0, x_i)\mathbf{W}_{0,rs} l(y_s^0,\cdot)\right\rangle_{\mathcal{H}} \\
&\quad - \frac{2}{n}\sum_{i=1}^n \left\langle \sum_{p,q=1}^{n_0} k_0(x_p^0, x_i)\mathbf{W}_{0,pq} l(y_q^0,\cdot), \sum_{r,s=1}^{n_1} k_1(x_r^1, x_i)\mathbf{W}_{1,rs} l(y_s^1,\cdot)\right\rangle_{\mathcal{H}} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \left\langle \sum_{p,q=1}^{n_1} k_1(x_p^1, x_i)\mathbf{W}_{1,pq} l(y_q^1,\cdot), \sum_{r,s=1}^{n_1} k_1(x_r^1, x_i)\mathbf{W}_{1,rs} l(y_s^1,\cdot)\right\rangle_{\mathcal{H}} \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{p,q,r,s=1}^{n_0} k_0(x_i, x_p^0)\mathbf{W}_{0,pq} l(y_q^0, y_s^0)\mathbf{W}_{0,sr}^T k_0(x_r^0, x_i) \\
&\quad - \frac{2}{n}\sum_{i=1}^n \sum_{p,q=1}^{n_0}\sum_{r,s=1}^{n_1} k_0(x_i, x_p^0)\mathbf{W}_{0,pq} l(y_q^0, y_s^1)\mathbf{W}_{1,sr}^T k_1(x_r^1, x_i) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \sum_{p,q,r,s=1}^{n_1} k_1(x_i, x_p^1)\mathbf{W}_{1,pq} l(y_q^1, y_s^1)\mathbf{W}_{1,sr}^T k_1(x_r^1, x_i)
\end{aligned}$$

$$= \frac{1}{n}\left\{\text{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\tilde{\mathbf{K}}_0^T\right) - 2\text{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right) + \text{Tr}\left(\tilde{\mathbf{K}}_1\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right)\right\}$$

□

Theorem 4.5. *Under the same assumptions as in Theorem 4.2, we have* $\hat{t} \xrightarrow{p} t$ *as* $n_0, n_1 \to \infty$.

Proof. We decompose $\left|\hat{t} - t\right|$ as follows using the triangle inequality:

$$\begin{aligned}
\left|\hat{t} - t\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2 - \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|_{\mathcal{H}}^2 - \mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right| \\
&\quad + \left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right] - \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right|
\end{aligned}$$

Here, the first term converges to 0 in probability by the uniform law of large numbers. For the second term, see that

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right] - \mathbb{E}\left[\left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right| \\
&= \left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X} + \mu_{Y_1|X} - \mu_{Y_0|X} + \mu_{Y_0|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2 - \left\|\mu_{Y_1|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right]\right| \\
&= \left|\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2 + \left\|\mu_{Y_0|X} - \hat{\mu}_{Y_0|X}\right\|_{\mathcal{H}}^2\right] + 2\left\langle\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}, \mu_{Y_1|X} - \mu_{Y_0|X}\right\rangle_{\mathcal{H}}\right. \\
&\quad \left. + 2\left\langle\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}, \mu_{Y_1|X} - \mu_{Y_0|X}\right\rangle_{\mathcal{H}} + 2\left\langle\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}, \hat{\mu}_{Y_0|X} - \hat{\mu}_{Y_0|X}\right\rangle_{\mathcal{H}}\right|.
\end{aligned}$$

Here, we have

$$\mathbb{E}\left[\left\|\hat{\mu}_{Y_1|X} - \mu_{Y_1|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0 \qquad \text{and} \qquad \mathbb{E}\left[\left\|\hat{\mu}_{Y_0|X} - \mu_{Y_0|X}\right\|_{\mathcal{H}}^2\right] \xrightarrow{p} 0$$

as in the proof of Theorem 4.2, so we are done. □

Theorem 5.1. *The solution* $\hat{F}_0$ *to the problem in (4) is*

$$\hat{F}_0\left(x_1, ..., x_r\right) = \sum_{i_1,...,i_r=1}^{n_0} k_0\left(x_{i_1}^0, x_1\right)...k_0\left(x_{i_r}^0, x_r\right)c_{i_1,...,i_r}$$

where the coefficients $c_{i_1,...,i_r} \in \mathbb{R}$ *are the unique solution of the* n^r *linear equations*

$$\sum_{j_1,...,j_r=1}^{n_0}\left(k_0\left(x_{i_1}^0, x_{j_1}^0\right)...k_0\left(x_{i_r}^0, x_{j_r}^0\right) + \binom{n_0}{r}\lambda_{n_0}^0\delta_{i_1j_1}...\delta_{i_rj_r}\right)c_{j_1,...,j_r} = h\left(y_{i_1}^0, ..., y_{i_r}^0\right).$$

Proof. Recall from (4) that

$$\hat{F}_0 = \underset{F\in\mathcal{H}_0^r}{\arg\min}\left\{\frac{1}{\binom{n_0}{r}}\sum\left(F\left(x_{i_1}^0, ..., x_{i_r}^0\right) - h\left(y_{i_1}^0, ..., y_{i_r}^0\right)\right)^2 + \lambda_{n_0}^0\left\|F\right\|_{\mathcal{H}_0^r}^2\right\},$$

where the summation is over the $\binom{n_0}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $1, ..., n_0$. Write

$$\hat{F}_0'\left(x_1, ..., x_r\right) = \sum_{i_1,...,i_r=1}^{n_0} k_0\left(x_{i_1}^0, x_1\right)...k_0\left(x_{i_r}^0, x_r\right)c_{i_1,...,i_r}$$

where the coefficients $c_{i_1,...,i_r} \in \mathbb{R}$ are the unique solution of the n^r linear equations

$$\sum_{j_1,...,j_r=1}^{n_0} \left(k_0\left(x^0_{i_1}, x^0_{j_1}\right)...k_0\left(x^0_{i_r}, x^0_{j_r}\right) + \binom{n_0}{r}\lambda^0_{n_0}\delta_{i_1 j_1}...\delta_{i_r j_r} \right) c_{j_1,...,j_r} = h\left(y^0_{i_1},...,y^0_{i_r}\right).$$

Also, for any $F \in \mathcal{H}^r_0$, write $\hat{\mathcal{E}}_{\text{reg}}(F)$ for the empirical regularised least-squares risk of F:

$$\hat{\mathcal{E}}_{\text{reg}}(F) = \frac{1}{\binom{n_0}{r}} \sum \left(F\left(x^0_{i_1},...,x^0_{i_r}\right) - h\left(y^0_{i_1},...,y^0_{i_r}\right)\right)^2 + \lambda^0_{n_0}\left\|F\right\|^2_{\mathcal{H}^r_0},$$

so that $\hat{F}_0 = \arg\min_{F\in\mathcal{H}^r_0} \hat{\mathcal{E}}_{\text{reg}}(F)$. We will show that $\hat{F}'_0 = \hat{F}_0$. For any $F \in \mathcal{H}^r_0$, write $G = F - \hat{F}'_0$. Then

$$\begin{aligned}
\hat{\mathcal{E}}_{\text{reg}}(F) &= \frac{1}{\binom{n_0}{r}} \sum \left(F\left(x^0_{i_1},...,x^0_{i_r}\right) - h\left(y^0_{i_1},...,y^0_{i_r}\right)\right)^2 + \lambda^0_{n_0}\left\|F\right\|^2_{\mathcal{H}^r_0} \\
&= \frac{1}{\binom{n_0}{r}} \sum \left(F\left(x^0_{i_1},...,x^0_{i_r}\right) - \hat{F}'_0\left(x^0_{i_1},...,x^0_{i_r}\right) + \hat{F}'_0\left(x^0_{i_1},...,x^0_{i_r}\right) - h\left(y^0_{i_1},...,y^0_{i_r}\right)\right)^2 + \lambda^0_{n_0}\left\|F\right\|^2_{\mathcal{H}^r_0} \\
&= \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}'_0\right) + \frac{1}{\binom{n_0}{r}} \sum G\left(x^0_{i_1},...,x^0_{i_r}\right)^2 + \frac{2}{\binom{n_0}{r}} \sum G\left(x^0_{i_1},...,x^0_{i_r}\right)\left(\hat{F}'_0\left(x^0_{i_1},...,x^0_{i_r}\right) - h\left(y^0_{i_1},...,y^0_{i_r}\right)\right) \\
&\quad + \lambda^0_{n_0}\left\|G\right\|^2_{\mathcal{H}^r_0} + 2\lambda^0_{n_0}\left\langle G, \hat{F}'_0\right\rangle_{\mathcal{H}^r_0} \\
&\geq \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}'_0\right) - \frac{2}{\binom{n_0}{r}} \sum G\left(x^0_{i_1},...,x^0_{i_r}\right)\left(h\left(y^0_{i_1},...,y^0_{i_r}\right) - \hat{F}'_0\left(x^0_{i_1},...,x^0_{i_r}\right)\right) + 2\lambda^0_{n_0}\left\langle G, \hat{F}'_0\right\rangle_{\mathcal{H}^r_0} \\
&= \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}'_0\right) - 2\lambda^0_{n_0} \sum G\left(x^0_{i_1},...,x^0_{i_r}\right) c_{i_1,...,i_r} + 2\lambda^0_{n_0} \sum_{i_1,...,i_r=1}^{n_0} G\left(x^0_{i_1},...,x^0_{i_r}\right) c_{i_1,...,i_r} \\
&\qquad \text{by the reproducing property and the definition of } c_{i_1,...,i_r} \\
&= \hat{\mathcal{E}}_{\text{reg}}\left(\hat{F}'_0\right)
\end{aligned}$$

Hence, $\hat{F}'_0$ minimises $\hat{\mathcal{E}}_{\text{reg}}$ in $\mathcal{H}^r_0$, and so $\hat{F}'_0 = \hat{F}_0$ as required. □

Theorem 5.2. *Suppose k^r_0 is a bounded and universal kernel and that $\lambda^0_{n_0}$ decays at a slower rate than $\mathcal{O}(n_0^{-1/2})$. Then as $n_0 \to \infty$,*

$$\mathbb{E}\left[\left(\hat{F}_0\left(X_1,...,X_r\right) - F_0\left(X_1,...,X_r\right)\right)^2\right] \xrightarrow{p} 0.$$

Proof. Define

$$F_{0,\lambda^0_{n_0}} = \arg\min_{F\in\mathcal{H}^r_0}\left\{\mathbb{E}\left[\left(F\left(X_1,...,X_r\right) - F_0\left(X_1,...,X_r\right)\right)^2\right] + \lambda^0_{n_0}\left\|F\right\|^2_{\mathcal{H}^r_0}\right\}.$$

By the bias-variance decomposition, this also minimises

$$\mathcal{E}_{\lambda^0_{n_0}}(F) = \mathbb{E}\left[\left(F\left(X_1,...,X_r\right) - h\left(Y_1,...,Y_r\right)\right)^2\right] + \lambda^0_{n_0}\left\|F\right\|^2_{\mathcal{H}^r_0}.$$

Denote the Hilbert space of P^r_X-square-integrable $\mathcal{X}^r \to \mathbb{R}$ functions by $L^2(\mathcal{X}^r, P^r_X)$, and define the inclusion operator

$$\iota : \mathcal{H}^r_0 \to L^2(\mathcal{X}^r, P^r_X).$$

Then we see that

$$\begin{aligned}
F_{0,\lambda^0_{n_0}} &= \arg\min_{F\in\mathcal{H}^r_0}\left\{\left\|\iota(F) - F_0\right\|^2_2 + \lambda^0_{n_0}\left\|F\right\|^2_{\mathcal{H}^r_0}\right\} \\
\implies \qquad 0 &= \iota^*(\iota(F_{0,\lambda^0_{n_0}}) - F_0) + \lambda^0_{n_0}F_{0,\lambda^0_{n_0}}
\end{aligned}$$

$$\implies \qquad F_{0,\lambda_{n_0}^0} = \left(\iota^* \circ \iota + \lambda_{n_0}^0 I\right)^{-1} \iota^* F_0$$

Now, for any $\mathbf{x}^0 = (x_1^0, ..., x_{n_0}^0)^T \in \mathcal{X}^{n_0}$, define the sampling operator

$$S_{\mathbf{x}^0} : \mathcal{H}_0^r \to \mathbb{R}^{\binom{n_0}{r}}, \qquad \left(S_{\mathbf{x}^0}(F)\right)_{i_1,...,i_r} = \frac{1}{\binom{n_0}{r}} F\left(x_{i_1}^0, ..., x_{i_r}^0\right), \{i_1, ..., i_r\} \subset \{1, ..., n_0\},$$

with adjoint

$$S_{\mathbf{x}^0}^* (\mathbf{h}) = \frac{1}{\binom{n_0}{r}} \sum k_0\left(x_{i_1}^0, \cdot\right) ... k_0\left(x_{i_r}^0, \cdot\right) h_{i_1,...,i_r}, \qquad \mathbf{h} \in \mathbb{R}^{\binom{n_0}{r}};$$

indeed, for any $F \in \mathcal{H}_0^r$ and $\mathbf{h} \in \mathbb{R}^{\binom{n_0}{r}}$,

$$\begin{aligned}
\langle S_{\mathbf{x}^0} F, \mathbf{h} \rangle_{\mathbb{R}^{\binom{n_0}{r}}} &= \frac{1}{\binom{n_0}{r}} \sum F\left(x_{i_1}^0, ..., x_{i_r}^0\right) h_{i_1,...,i_r} \\
&= \frac{1}{\binom{n_0}{r}} \sum \left\langle F, k_0\left(x_{i_1}^0, \cdot\right) ... k_0\left(x_{i_r}^0, \cdot\right) \right\rangle_{\mathcal{H}_0^r} h_{i_1,...,i_r} \\
&= \left\langle F, \frac{1}{\binom{n_0}{r}} \sum k_0\left(x_{i_1}^0, \cdot\right) ... k_0\left(x_{i_r}^0, \cdot\right) h_{i_1,...,i_r} \right\rangle_{\mathcal{H}_0^r}.
\end{aligned}$$

For $\mathbf{y}^0 \in \mathcal{Y}^{n_0}$, write

$$h\left(\mathbf{y}^0\right) \in \mathbb{R}^{\binom{n_0}{r}}, \qquad h\left(\mathbf{y}^0\right)_{i_1,...,i_r} = h\left(y_{i_1}^0, ..., y_{i_r}^0\right), \{i_1, ..., i_r\} \subset \{1, ..., n_0\}.$$

Then we see that

$$\begin{aligned}
\hat{F}_0 &= \underset{F \in \mathcal{H}_0^r}{\arg\min} \left\{ \binom{n_0}{r} \left\| S_{\mathbf{x}^0}(F) - \frac{1}{\binom{n_0}{r}} h\left(\mathbf{y}^0\right) \right\|^2 + \lambda_{n_0}^0 \|F\|_{\mathcal{H}_0^r}^2 \right\} \\
\implies \qquad 0 &= \binom{n_0}{r} S_{\mathbf{x}^0}^* \left(S_{\mathbf{x}^0}\left(\hat{F}_0\right) - \frac{1}{\binom{n_0}{r}} h\left(\mathbf{y}^0\right) \right) + \lambda_{n_0}^0 \hat{F}_0 \\
\implies \qquad \hat{F}_0 &= \left(\binom{n_0}{r} S_{\mathbf{x}^0}^* \circ S_{\mathbf{x}^0} + \lambda_{n_0}^0 I \right)^{-1} S_{\mathbf{x}^0}^* h\left(\mathbf{y}^0\right).
\end{aligned}$$

We consider the following decomposition:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{F}_0\left(X_1, ..., X_r\right) - F_0\left(X_1, ..., X_r\right) \right)^2 \right] = \left\| \iota \hat{F}_0 - F_0 \right\|_2^2 &\leq 2 \left\| \iota \hat{F}_0 - \iota F_{0,\lambda_{n_0}^0} \right\|_2^2 \qquad \text{(a)} \\
&+ 2 \left\| \iota F_{0,\lambda_{n_0}^0} - F_0 \right\|_2^2. \qquad \text{(b)}
\end{aligned}$$

We are done if we show that the terms (a) and (b) separately converge to 0 (in probability, for (a)).

(a) See that

$$\begin{aligned}
\hat{F}_0 - F_{0,\lambda_{n_0}^0} &= \left(\binom{n_0}{r} S_{\mathbf{x}^0}^* \circ S_{\mathbf{x}^0} + \lambda_{n_0}^0 I \right)^{-1} S_{\mathbf{x}^0}^* h\left(\mathbf{y}^0\right) - F_{0,\lambda_{n_0}^0} \\
&= \left(\binom{n_0}{r} S_{\mathbf{x}^0}^* \circ S_{\mathbf{x}^0} + \lambda_{n_0}^0 I \right)^{-1} \left(S_{\mathbf{x}^0}^* h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S_{\mathbf{x}^0}^* \circ S_{\mathbf{x}^0} F_{0,\lambda_{n_0}^0} + \iota^* \left(\iota F_{0,\lambda_{n_0}^0} - F_0 \right) \right).
\end{aligned}$$

By spectral theorem,

$$\left\|\hat{F}_0 - F_{0,\lambda^0_{n_0}}\right\|_{\mathcal{H}} \leq \frac{1}{\lambda^0_{n_0}} \left\| S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} + \iota^*\left(\iota F_{0,\lambda^0_{n_0}} - F_0\right)\right\|_{\mathcal{H}}.$$

Using this inequality and Chebyshev's inequality, for any $\epsilon > 0$,

$$\begin{aligned}
&P\left(\left\|\hat{F}_0 - F_{0,\lambda^0_{n_0}}\right\|_{\mathcal{H}} \geq \epsilon\right) \leq P\left(\frac{1}{\lambda^0_{n_0}} \left\| S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} - \iota^*\left(F_0 - \iota F_{0,\lambda^0_{n_0}}\right)\right\|_{\mathcal{H}} \geq \epsilon\right) \\
&\leq \frac{1}{(\lambda^0_{n_0})^2 \epsilon^2} \mathbb{E}\left[\left\| S^*_{\mathbf{x}^0} h\left(\mathbf{y}^0\right) - \binom{n_0}{r} S^*_{\mathbf{x}^0} \circ S_{\mathbf{x}^0} F_{0,\lambda^0_{n_0}} - \iota^*\left(F_0 - \iota F_{0,\lambda^0_{n_0}}\right)\right\|^2_{\mathcal{H}}\right] \\
&\leq \frac{1}{(\lambda^0_{n_0})^2 \epsilon^2 \binom{n_0}{r}} \mathbb{E}\left[\left\| k_0\left(x^0_{i_1}, \cdot\right) ... k_0\left(x^0_{i_r}, \cdot\right)\left(h\left(y^0_{i_1}, ..., y^0_{i_r}\right) - F_{0,\lambda^0_{n_0}}\left(x^0_{i_1}, ..., x^0_{i_r}\right)\right)\right\|^2_{\mathcal{H}}\right] \\
&\to 0
\end{aligned}$$

as $n \to \infty$, since the kernel is bounded.

(b) Take an arbitrary $\epsilon > 0$. By the denseness of $\mathcal{H}^r_0$ in $L^2(\mathcal{X}^r, P^r_X)$, there exists some $F_\epsilon \in \mathcal{H}^r_0$ with

$$\|\iota F_\epsilon - F_0\|^2_2 = \mathcal{E}(F_\epsilon) - \mathcal{E}(F_0) \leq \frac{\epsilon}{2}.$$

Then

$$\begin{aligned}
\left\|\iota F_{0,\lambda^0_{n_0}} - F_0\right\|^2_2 &= \mathcal{E}\left(F_{0,\lambda^0_{n_0}}\right) - \mathcal{E}\left(F_0\right) \\
&\leq \mathcal{E}_{\lambda^0_{n_0}}\left(F_{0,\lambda^0_{n_0}}\right) - \mathcal{E}\left(F_0\right) \\
&= \mathcal{E}_{\lambda^0_{n_0}}\left(F_{0,\lambda^0_{n_0}}\right) - \mathcal{E}_{\lambda^0_{n_0}}(F_\epsilon) + \mathcal{E}_{\lambda^0_{n_0}}(F_\epsilon) - \mathcal{E}(F_\epsilon) + \mathcal{E}(F_\epsilon) - \mathcal{E}\left(F_0\right) \\
&\leq \lambda^0_{n_0} \|F_\epsilon\|^2_{\mathcal{H}^r_0} + \frac{\epsilon}{2}.
\end{aligned}$$

Now let n be large enough for

$$\lambda^0_{n_0} \|F_\epsilon\|^2_{\mathcal{H}^r_0} \leq \frac{\epsilon}{2}$$

to hold.

□