A. Topological and Measurability Considerations

Let $(\mathcal{Z}, \tau_\mathcal{Z})$ and $(\mathcal{L}, \tau_\mathcal{L})$ be two separable topological spaces, where $\mathcal{Z}$ is the *input space* and $\mathcal{L} := \{-1, 1\}$ is the *output space*. Let $\mathcal{B}(\tau)$ be the Borel σ-algebra induced by the topology τ. Let $\mathbb{P}$ be an unknown probability measure on $(\mathcal{Z} \times \mathcal{L}, \mathcal{B}(\tau_\mathcal{Z}) \otimes \mathcal{B}(\tau_\mathcal{L}))$.

Consider also the classifiers $f \in \mathcal{F}_k$ and loss function ℓ to be measurable.

A.1. Measurability Conditions to Learn from Distributions

The first step towards the deployment of our learning setup is to guarantee the existence of a measure on the space $\mu_k(\mathcal{P}) \times \mathcal{L}$, where $\mu_k(\mathcal{P}) = \{\mu_k(P) : P \in \mathcal{P}\} \subseteq \mathcal{H}_k$ is the set of kernel mean embeddings associated with the measures in $\mathcal{P}$. The following lemma provides such guarantee. This allows the analysis within the rest of this Section on $\mu_k(\mathcal{P}) \times \mathcal{L}$.

Lemma 2. *Let $(\mathcal{Z}, \tau_\mathcal{Z})$ and $(\mathcal{L}, \tau_\mathcal{L})$ be two separable topological spaces. Let $\mathcal{P}$ be the set of all Borel probability measures on $(\mathcal{Z}, \mathcal{B}(\tau_\mathcal{Z}))$. Let $\mu_k(\mathcal{P}) = \{\mu_k(P) : P \in \mathcal{P}\} \subseteq \mathcal{H}_k$, where μ_k is the kernel mean embedding* (1) *associated to some bounded continuous kernel function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Then, there exists a measure on $\mu_k(\mathcal{P}) \times \mathcal{L}$.*

Proof. The following is a similar result to Szabó et al. (2014, Proof 3).

Start by endowing $\mathcal{P}$ with the weak topology $\tau_\mathcal{P}$, such that the map

$$L(P) = \int_\mathcal{Z} f(z) \mathrm{d}P(z), \tag{17}$$

is continuous for all $f \in C_b(\mathcal{Z})$. This makes $(\mathcal{P}, \mathcal{B}(\tau_\mathcal{P}))$ a measurable space.

First, we show that $\mu_k : (\mathcal{P}, \mathcal{B}(\tau_\mathcal{P})) \to (\mathcal{H}_k, \mathcal{B}(\tau_\mathcal{H}))$ is Borel measurable. Note that $\mathcal{H}_k$ is separable due to the separability of $(\mathcal{Z}, \tau_\mathcal{Z})$ and the continuity of k (Steinwart & Christmann, 2008, Lemma 4.33). The separability of $\mathcal{H}_k$ implies μ_k is Borel measurable iff it is weakly measurable (Reed & Simon, 1972, Thm. IV.22). Note that the boundedness and the continuity of k imply $\mathcal{H}_k \subseteq C_b(\mathcal{Z})$ (Steinwart & Christmann, 2008, Lemma 4.28). Therefore, (17) remains continuous for all $f \in \mathcal{H}_k$, which implies the Borel measurability of μ_k.

Second, $\mu_k : (\mathcal{P}, \mathcal{B}(\tau_\mathcal{P})) \to (\mathcal{G}, \mathcal{B}(\tau_\mathcal{G}))$ is Borel measurable, since the $\mathcal{B}(\tau_\mathcal{G}) = \{A \cap \mathcal{G} : A \in \mathcal{B}(\mathcal{H}_k)\} \subseteq \mathcal{B}(\tau_\mathcal{H})$, where $\mathcal{B}(\tau_\mathcal{G})$ is the σ-algebra induced by the topology of $\mathcal{G} \in \mathcal{B}(\mathcal{H}_k)$ (Szabó et al., 2014).

Third, we show that $g : (\mathcal{P} \times \mathcal{L}, \mathcal{B}(\tau_\mathcal{P}) \otimes \mathcal{B}(\tau_\mathcal{L})) \to (\mathcal{G} \times \mathcal{L}, \mathcal{B}(\tau_\mathcal{G}) \otimes \mathcal{B}(\tau_\mathcal{L}))$ is measurable. For that, it suffices to decompose $g(x, y) = (g_1(x, y), g_2(x, y))$ and show that g_1 and g_2 are measurable (Szabó et al., 2014). □

B. Proofs

B.1. Theorem 1

Note that the original statement of Theorem 27 in Song (2008) assumed $f \in [0, 1]$ while we let elements of the ball in RKHS to take negative values as well which can be achieved by minor changes of the proof. For completeness we provide the modified proof here. Using the well known dual relation between the norm in RKHS and sup-norm of empirical process which can be found in Theorem 28 of Song (2008) we can write:

$$\|\mu_k(P) - \mu_k(P_S)\|_{\mathcal{H}_k} = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \left(\mathop{\mathbb{E}}_{z \sim P}[f(z)] - \frac{1}{n} \sum_{i=1}^{n} f(z_i) \right). \tag{18}$$

Now we proceed in the usual way. First we note that the sup-norm of empirical process appearing on the r.h.s. can be viewed as a real-valued function of i.i.d. random variables $z_1, \ldots, z_n$. We will denote it as $F(z_1, \ldots, z_n)$. The straightforward computations show that the function F satisfies the *bounded difference* condition (Theorem 14 of Song (2008)). Indeed, let us fix all the values $z_1, \ldots, z_n$ except for the z_j which we will set to z_j'. Using identity $|a - b| = (a - b)\mathbb{1}_{a>b} + (b - a)\mathbb{1}_{a \leq b}$ and noting that if $\sup_x f(x) = f(x^*)$ then $\sup_x f(x) - \sup_x g(x)$ is upper bounded by $f(x^*) - g(x^*)$ we get

$$\begin{aligned}
&|F(z_1, \ldots, z_j', \ldots, z_n) - F(z_1, \ldots, z_j, \ldots, z_n)| \\
&\leq \frac{1}{n}\big(f(z_j) - f(z_j')\big)\mathbb{1}_{F(z_1,\ldots,z_j',\ldots,z_n) > F(z_1,\ldots,z_j,\ldots,z_n)} + \frac{1}{n}\big(f(z_j') - f(z_j)\big)\mathbb{1}_{F(z_1,\ldots,z_j',\ldots,z_n) \leq F(z_1,\ldots,z_j,\ldots,z_n)}.
\end{aligned}$$

Now noting that $|f(z) - f(z')| \in [0,2]$ we conclude with

$$\begin{aligned}&|F(z_1,\dots,z_j',\dots,z_n) - F(z_1,\dots,z_j,\dots,z_n)|\\ &\leq \frac{2}{n}\mathbb{1}_{F(z_1,\dots,z_j',\dots,z_n)>F(z_1,\dots,z_j,\dots,z_n)} + \frac{2}{n}\mathbb{1}_{F(z_1,\dots,z_j',\dots,z_n)\leq F(z_1,\dots,z_j,\dots,z_n)} = \frac{2}{n}.\end{aligned}$$

Using McDiarmid's inequality (Theorem 14 of (Song, 2008)) with $c_i = 2/n$ we obtain that with probability not less than $1-\delta$ the following holds:

$$\sup_{\|f\|_{\mathcal{H}_k}\leq 1}\left(\mathop{\mathbb{E}}_{z\sim P}[f(z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\right) \leq \mathbb{E}\left[\sup_{\|f\|_{\mathcal{H}_k}\leq 1}\left(\mathop{\mathbb{E}}_{z\sim P}[f(z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\right)\right] + \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Finally, we proceed with the symmetrization step (Theorem 2.1 of (Koltchinskii, 2011)) which upper bounds the expected value of the sup-norm of empirical process with twice the Rademacher complexity of the class $\{f \in \mathcal{H}_k \colon \|f\|_{\mathcal{H}_k} \leq 1\}$ and with upper bound on this Rademacher complexity which can be found in Lemma 22 and related remarks of Bartlett & Mendelson (2002).

We also note that the original statement of Theorem 27 in Song (2008) contains extra factor of 2 under logarithm compared to our modified result. This is explained by the fact that while we upper bounded the Rademacher complexity directly, Song (2008) instead upper bounds it in terms of the empirical (or *conditional*) Rademacher complexity which results in another application of McDiarmid's inequality together with union bound.

B.2. Theorem 3

We will proceed as follows:

$$\begin{aligned}R_\varphi(\tilde f_n) - R_\varphi(f^*) &= R_\varphi(\tilde f_n) - \tilde R_\varphi(\tilde f_n)\\ &\quad + \tilde R_\varphi(\tilde f_n) - \tilde R_\varphi(f^*)\\ &\quad + \tilde R_\varphi(f^*) - R_\varphi(f^*)\\ &\leq 2\sup_{f\in\mathcal{F}_k}|R_\varphi(f) - \tilde R_\varphi(f)|\\ &= 2\sup_{f\in\mathcal{F}_k}|R_\varphi(f) - \hat R_\varphi(f) + \hat R_\varphi(f) - \tilde R_\varphi(f)|\\ &\leq 2\sup_{f\in\mathcal{F}_k}|R_\varphi(f) - \hat R_\varphi(f)| + 2\sup_{f\in\mathcal{F}_k}|\hat R_\varphi(f) - \tilde R_\varphi(f)|. \qquad (19)\end{aligned}$$

We will now upper bound two terms in (19) separately.

We start with noticing that Theorem 2 can be used in order to upper bound the first term. All we need is to match the quantities appearing in our problem to the classical setting of learning theory, discussed in Section 3.1. Indeed, let $\mu(\mathcal{P})$ play the role of input space $\mathcal{Z}$. Thus the input objects are kernel mean embeddings of elements of $\mathcal{P}$. According to Lemma 2, there is a distribution defined over $\mu(\mathcal{P})\times\mathcal{L}$, which will play the role of unknown distribution $\mathbb{P}$. Finally, i.i.d. pairs $\left\{\left(\mu_k(P_i), l_i\right)\right\}_{i=1}^n$ form the training sample. Thus, using Theorem 2 we get that with probability not less than $1-\delta/2$ (w.r.t. the random training sample $\left\{\left(\mu_k(P_i), l_i\right)\right\}_{i=1}^n$) the following holds true:

$$\sup_{f\in\mathcal{F}_k}|R_\varphi(f) - \hat R_\varphi(f)| \leq 2L_\varphi\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_k}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i f(z_i)\right|\right] + B\sqrt{\frac{\log(2/\delta)}{2n}}. \qquad (20)$$

To deal with the second term in (19) we note that

$$\begin{aligned}\sup_{f\in\mathcal{F}_k}|\hat R_\varphi(f) - \tilde R_\varphi(f)| &= \sup_{f\in\mathcal{F}_k}\left|\frac{1}{n}\sum_{i=1}^n\Big[\varphi\big(-l_i f\big(\mu_k(P_i)\big)\big) - \varphi\big(-l_i f\big(\mu_k(P_{S_i})\big)\big)\Big]\right|\\ &\leq \sup_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n\left|\varphi\big(-l_i f\big(\mu_k(P_i)\big)\big) - \varphi\big(-l_i f\big(\mu_k(P_{S_i})\big)\big)\right|\\ &\leq L_\varphi\sup_{f\in\mathcal{F}_k}\frac{1}{n}\sum_{i=1}^n\left|f\big(\mu_k(P_i)\big) - f\big(\mu_k(P_{S_i})\big)\right|,\end{aligned}$$

where we have used the Lipschitzness of the cost function φ. Using the Lipschitzness of the functionals $f \in \mathcal{F}_k$ we obtain:

$$\sup_{f\in\mathcal{F}_k} |\hat{R}_\varphi(f) - \tilde{R}_\varphi(f)| \leq L_\varphi \sup_{f\in\mathcal{F}_k} \frac{L_f}{n}\sum_{i=1}^n \|\mu_k(P_i) - \mu_k(P_{S_i})\|_{\mathcal{H}_k}. \tag{21}$$

Also note that the usual reasoning shows that if $h \in \mathcal{H}_k$ and $\|h\|_{\mathcal{H}_k} \leq 1$ then:

$$|h(z)| = |\langle h, k(z,\cdot)\rangle_{\mathcal{H}_k}| \leq \|h\|_{\mathcal{H}_k}\|k(z,\cdot)\|_{\mathcal{H}_k} = \|h\|_{\mathcal{H}_k}\sqrt{k(z,z)} \leq \sqrt{k(z,z)}$$

and hence $\|h\|_\infty = \sup_{z\in\mathcal{Z}}|h(z)| \leq 1$ because our kernel is bounded. This allows us to use Theorem 1 to control every term in (21) and combine the resulting upper bounds in a union bound[4] over $i = 1, \ldots, n$ to show that for any fixed $P_1, \ldots, P_n$ with probability not less than $1 - \delta/2$ (w.r.t. the random samples $\{S_i\}_{i=1}^n$) the following is true:

$$L_\varphi \sup_{f\in\mathcal{F}} \frac{L_f}{n}\sum_{i=1}^n \|\mu_k(P_i) - \mu_k(P_{S_i})\|_{\mathcal{H}_k} \leq L_\varphi \sup_{f\in\mathcal{F}} \frac{L_f}{n}\sum_{i=1}^n \left(2\sqrt{\frac{\mathbb{E}_{z\sim P}[k(z,z)]}{n_i}} + \sqrt{\frac{2\log\frac{2n}{\delta}}{n_i}}\right). \tag{22}$$

The quantity $2n/\delta$ appears under the logarithm since for every i we have used Theorem 1 with $\delta' = \delta/(2n)$. Combining (20) and (22) in a union bound together with (19) we finally get that with probability not less than $1 - \delta$ the following is true:

$$R_\varphi(\tilde{f}_n) - R_\varphi(f^*) \leq 4L_\varphi R_n(\mathcal{F}) + 2B\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{4L_\varphi L_{\mathcal{F}}}{n}\sum_{i=1}^n\left(\sqrt{\frac{\mathbb{E}_{z\sim P}[k(z,z)]}{n_i}} + \sqrt{\frac{\log\frac{2n}{\delta}}{2n_i}}\right),$$

where we have defined $L_{\mathcal{F}} = \sup_{f\in\mathcal{F}} L_f$.

B.3. Theorem 4

Our proof is a simple combination of the duality equation (18) combined with the following lower bound on the supremum of empirical process presented in Theorem 2.3 of Bartlett & Mendelson (2006):

Theorem 5. *Let F be a class of real-valued functions defined on a set $\mathcal{Z}$ such that $\sup_{f\in F}\|f\|_\infty \leq 1$. Let $z_1, \ldots, z_n, z \in \mathcal{Z}$ be i.i.d. according to some probability measure P on $\mathcal{Z}$. Set $\sigma_F^2 = \sup_{f\in F}\mathbb{V}[f(z)]$. Then there are universal constants c, c', and C for which the following holds:*

$$\mathbb{E}\left[\sup_{f\in F}\left|\mathbb{E}[f(z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\right|\right] \geq c\frac{\sigma_F}{\sqrt{n}}.$$

Furthermore, for every integer $n \geq 1/\sigma_F^2$, with probability at least c',

$$\sup_{f\in F}\left|\mathbb{E}[f(z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\right| \geq C\,\mathbb{E}\left[\sup_{f\in F}\left|\mathbb{E}[f(z)] - \frac{1}{n}\sum_{i=1}^n f(z_i)\right|\right].$$

We note that constants c, c', and C appearing in the last result do not depend on n, σ_F^2 or any other quantities appearing in the statement. This can be verified by the inspection of the proof presented in (Bartlett & Mendelson, 2006).

B.4. Lemma 1

Proof. Bochner's theorem (Rudin, 1962) states that for any shift-invariant symmetric p.d. kernel k defined on $\mathcal{Z}\times\mathcal{Z}$ where $\mathcal{Z} = R^d$ and any $z, z' \in \mathcal{Z}$ the following holds:

$$k(z,z') = \int_{\mathcal{Z}} p_k(w)e^{i\langle w, z-z'\rangle}dw, \tag{23}$$

[4] Note that the union bound results in the extra $\log n$ factor in our bound. We believe that this factor can be avoided using a refined proof technique, based on the application of McDiarmid's inequality. This question is left for a future work.

where p_k is a positive and integrable Fourier transform of the kernel k. It is immediate to check that Fourier transform of such kernels k is always an even function, meaning $p_k(-w) = p_k(w)$. Indeed, since $k(z-z') = k(z'-z)$ for all $z, z' \in \mathcal{Z}$ (due to symmetry of the kernel) we have:

$$p_k(w) := \int_{\mathcal{Z}} k(\delta)e^{i\langle w,\delta\rangle}d\delta = \int_{\mathcal{Z}} k(\delta)\cos(\langle w,\delta\rangle)d\delta = \int_{\mathcal{Z}} k(\delta)\cos(-\langle w,\delta\rangle)d\delta = p_k(-w)$$

which holds for any $w \in \mathbb{R}^d$. Thus for any $z, z' \in \mathbb{R}^d$ we can write:

$$\begin{aligned}
k(z,z') &= \int_{\mathbb{R}^d} p_k(w)\big(\cos(\langle w, z-z'\rangle) + i\cdot\sin(\langle w, z-z'\rangle)\big)dw \\
&= \int_{\mathbb{R}^d} p_k(w)\big(\cos(\langle w, z-z'\rangle)dw + i\cdot\int_{\mathbb{R}^d} p_k(w)\sin(\langle w, z-z'\rangle)\big)dw \\
&= \int_{\mathbb{R}^d} p_k(w)\cos(\langle w, z-z'\rangle)dw \\
&= 2\int_{\mathbb{R}^d}\int_0^{2\pi}\frac{1}{2\pi}p_k(w)\cos(\langle w,z\rangle + b)\cos(\langle w,z'\rangle + b)\,db\,dw.
\end{aligned}$$

Denote $C_k = \int_{\mathbb{R}^d} p(w)dw < \infty$. Next we will use identity $\cos(a-b) = \frac{1}{\pi}\int_0^{2\pi}\cos(a+x)\cos(b+x)dx$ and introduce random variables b and w distributed according to $\mathcal{U}[0,2\pi]$ and $\frac{1}{C_k}p_k(w)$ respectively. Then we can rewrite

$$k(z,z') = 2C_k \mathop{\mathbb{E}}_{b,w}\left[\cos(\langle w,z\rangle + b)\cos(\langle w,z'\rangle + b)\right]. \tag{24}$$

Now let Q be any probability distribution defined on $\mathcal{Z}$. Then for any $z, w \in \mathcal{Z}$ and $b \in [0,2\pi]$ the function

$$g^z_{w,b}(\cdot) := 2C_k\cos(\langle w,z\rangle + b)\cos(\langle w,\cdot\rangle + b)$$

belongs to the $L_2(Q)$ space. Namely, $L_2(Q)$ norm of such a function is finite. Moreover, it is bounded by $2C_k$:

$$\begin{aligned}
\|g^z_{w,b}(\cdot)\|^2_{L_2(Q)} &= \int_{\mathcal{Z}}\Big(2C_k\cos(\langle w,z\rangle + b)\cos(\langle w,t\rangle + b)\Big)^2 dQ(t) \\
&\le 4C_k^2\int_{\mathcal{Z}} dQ(t) = 4C_k^2.
\end{aligned} \tag{25}$$

Note that for any fixed $x \in \mathcal{Z}$ and any random parameters $w \in \mathcal{Z}$ and $b \in [0,2\pi]$ the element $g^z_{w,b}(\cdot)$ is a *random variable* taking values in the $L_2(Q)$ space (which is Hilbert). Such Banach-space valued random variables are well studied objects (Ledoux & Talagrand, 1991) and a number of concentration results for them are known by now. We will use the following version of Hoeffding inequality which can be found in Lemma 4 of Rahimi & Recht (2008):

Lemma 3. *Let $v_1, \ldots, v_m$ be i.i.d. random variables taking values in a ball of radius M centred around origin in a Hilbert space H. Then, for any $\delta > 0$, the following holds:*

$$\left\|\frac{1}{m}\sum_{i=1}^m v_i - \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m v_i\right]\right\|_H \le \frac{M}{m}\left(1 + \sqrt{2\log(1/\delta)}\right).$$

with probability higher than $1-\delta$ over the random sample $v_1, \ldots, v_m$.

Note that Bochner's formula (23) and particularly its simplified form (24) indicates that if w is distributed according to normalized Fourier transform $\frac{1}{C_k}p_k$ and $b \sim \mathcal{U}([0,2\pi])$ then $\mathbb{E}_{w,b}[g^z_{w,b}(\cdot)] = k(z,\cdot)$. Moreover, we can show that any element h of RKHS $\mathcal{H}_k$ also belongs to the $L_2(Q)$ space:

$$\begin{aligned}
\|h(\cdot)\|^2_{L_2(Q)} &= \int_{\mathcal{Z}}\big(h(t)\big)^2 dQ(t) \\
&= \int_{\mathcal{Z}}\langle k(t,\cdot), h(\cdot)\rangle^2_{\mathcal{H}_k} dQ(t) \\
&\le \int_{\mathcal{Z}} k(t,t)\|h\|^2_{\mathcal{H}_k} dQ(t) \le \|h\|^2_{\mathcal{H}_k} < \infty,
\end{aligned} \tag{26}$$

where we have used the reproducing property of k in RKHS $\mathcal{H}_k$, Cauchy-Schwartz inequality, and the fact that the kernel k is bounded. Thus we conclude that the function $k(z,\cdot)$ is also an element of $L_2(Q)$ space.

This shows that if we have a sample of i.i.d. pairs $\{(w_i,b_i)\}_{i=1}^m$ then $\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m g^z_{w_i,b_i}(\cdot)\right] = k(z,\cdot)$ where $\{g^z_{w_i,b_i}(\cdot)\}_{i=1}^m$ are i.i.d. elements of Hilbert space $L_2(Q)$. We conclude the proof using concentration inequality for Hilbert spaces of Lemma 3 and a union bound over the elements $z \in S$, since

$$\begin{aligned}
\left\|\mu_k(P_S) - \frac{1}{n}\sum_{i=1}^n \hat{g}^{z_i}_m(\cdot)\right\|_{L_2(Q)} &= \left\|\frac{1}{n}\sum_{i=1}^n k(z_i,\cdot) - \frac{1}{n}\sum_{i=1}^n \hat{g}^{z_i}_m(\cdot)\right\|_{L_2(Q)}\\
&\leq \frac{1}{n}\sum_{i=1}^n \|k(z_i,\cdot) - \hat{g}^{z_i}_m(\cdot)\|_{L_2(Q)}\\
&= \frac{1}{n}\sum_{i=1}^n \left\|k(z_i,\cdot) - \frac{1}{m}\sum_{i=j}^m g^{z_i}_{w_j,b_j}(\cdot)\right\|_{L_2(Q)},
\end{aligned}$$

where we have used the triangle inequality. □

B.5. Excess Risk Bound for Low-Dimensional Representations

Let us first recall some important notations introduced in Section 3.3. For any $w, z \in \mathcal{Z}$ and $b \in [0, 2\pi]$ we define the following functions

$$g^z_{w,b}(\cdot) = 2C_k \cos(\langle w, z\rangle + b)\cos(\langle w, \cdot\rangle + b) \in L_2(Q), \tag{27}$$

where $C_k = \int_{\mathcal{Z}} p_k(z)dz$ for $p_k \colon \mathcal{Z} \to \mathbb{R}$ being the Fourier transform of k. We sample m pairs $\{(w_i,b_i)\}_{i=1}^m$ i.i.d. from $\left(\frac{1}{C_k}p_k\right) \times \mathcal{U}[0,2\pi]$ and define the average function

$$\hat{g}^z_m(\cdot) = \frac{1}{m}\sum_{i=1}^m g^z_{w_i,b_i}(\cdot) \in L_2(Q).$$

Since cosine functions (27) do not necessarily belong to the RKHS $\mathcal{H}_k$ and we are going to use their linear combinations as a training points, our classifiers should now act on the whole $L_2(Q)$ space. To this end, we redefine the set of classifiers introduced in the Section 3.2 to be $\{\text{sign} \circ f \colon f \in \mathcal{F}_Q\}$ where now $\mathcal{F}_Q$ is the set of functionals mapping $L_2(Q)$ to $\mathbb{R}$.

Recall that our goal is to find f^* such that

$$f^* \in \arg\min_{f \in \mathcal{F}_Q} R_\varphi(f) := \arg\min_{f \in \mathcal{F}_Q} \mathop{\mathbb{E}}_{(P,l)\sim\mathcal{M}}\left[\varphi\Big(-f\big(\mu_k(P)\big)l\Big)\right]. \tag{28}$$

As was pointed out in Section B.4 if the kernel k is bounded $\sup_{z\in\mathcal{Z}} k(z,z) \leq 1$ then $\mathcal{H}_k \subseteq L_2(Q)$. In particular, for any $P \in \mathcal{P}$ it holds that $\mu_k(P) \in L_2(Q)$ and thus (28) is well defined.

Instead of solving (28) directly, we will again use the version of empirical risk minimization (ERM). However, this time we won't use empirical mean embeddings $\{\mu_k(P_{S_i})\}_{i=1}^n$ since, as was already discussed, those lead to the expensive computations involving the kernel matrix. Instead, we will pose the ERM problem in terms of the low-dimensional approximations based on cosines. Namely, we propose to use the following estimator $\tilde{f}^m_n$:

$$\tilde{f}^m_n \in \arg\min_{f \in \mathcal{F}_Q} \tilde{R}^m_\varphi(f) := \arg\min_{f \in \mathcal{F}_Q} \frac{1}{n}\sum_{i=1}^n \varphi\left(-f\left(\frac{1}{n_i}\sum_{z\in S_i}\hat{g}^z_m(\cdot)\right)l_i\right).$$

The following result puts together Theorem 3 and Lemma 1 to provide an excess risk bound for $\tilde{f}^m_n$ which accounts for all sources of the errors introduced in the learning pipeline:

Theorem 6. *Let $\mathcal{Z} = \mathbb{R}^d$ and Q be any probability distribution on $\mathcal{Z}$. Consider the RKHS $\mathcal{H}_k$ associated with some bounded, continuous, shift-invariant kernel function k, such that* $\sup_{z\in\mathcal{Z}} k(z,z) \leq 1$. *Consider a class $\mathcal{F}_Q$ of functionals* *mapping $L_2(Q)$ to $\mathbb{R}$ with Lipschitz constants uniformly bounded by L_Q. Let $\varphi\colon \mathbb{R} \to \mathbb{R}^+$ be a L_φ-Lipschitz function such that $\phi(z) \geq \mathbb{1}_{z>0}$. Let $\varphi\bigl(-f(h)l\bigr) \leq B$ for every $f \in \mathcal{F}_Q$, $h \in L_2(Q)$, and $l \in \mathcal{L}$. Then for any $\delta > 0$ the following holds:*

$$\begin{aligned}
R_\varphi(\tilde{f}_n^m) - R_\varphi(f^*) &\leq 4L_\varphi R_n(\mathcal{F}_Q) + 2B\sqrt{\frac{\log(3/\delta)}{2n}}\\
&\quad + \frac{4L_\varphi L_Q}{n}\sum_{i=1}^n\left(\sqrt{\frac{\mathbb{E}_{z\sim P_i}[k(z,z)]}{n_i}} + \sqrt{\frac{\log\frac{3n}{\delta}}{2n_i}}\right)\\
&\quad + 2\frac{L_\varphi L_Q}{n}\sum_{i=1}^n \frac{2C_k}{\sqrt{m}}\left(1 + \sqrt{2\log(3n\cdot n_i/\delta)}\right)
\end{aligned}$$

with probability not less than $1-\delta$ over all sources of randomness, which are $\{(P_i,l_i)\}_{i=1}^n$, $\{S_i\}_{i=1}^n$, $\{(w_i,b_i)\}_{i=1}^m$.

Proof. We will proceed similarly to (19):

$$\begin{aligned}
R_\varphi(\tilde{f}_n^m) - R_\varphi(f^*) &= R_\varphi(\tilde{f}_n^m) - \tilde{R}_\varphi^m(\tilde{f}_n^m)\\
&\quad + \tilde{R}_\varphi^m(\tilde{f}_n^m) - \tilde{R}_\varphi^m(f^*)\\
&\quad + \tilde{R}_\varphi^m(f^*) - R_\varphi(f^*)\\
&\leq 2\sup_{f\in\mathcal{F}_Q}|R_\varphi(f) - \tilde{R}_\varphi^m(f)|\\
&= 2\sup_{f\in\mathcal{F}_Q}|R_\varphi(f) - \hat{R}_\varphi(f) + \hat{R}_\varphi(f) - \tilde{R}_\varphi(f) + \tilde{R}_\varphi(f) - \tilde{R}_\varphi^m(f)|\\
&\leq 2\sup_{f\in\mathcal{F}_Q}|R_\varphi(f) - \hat{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}_Q}|\hat{R}_\varphi(f) - \tilde{R}_\varphi(f)| + 2\sup_{f\in\mathcal{F}_Q}|\tilde{R}_\varphi(f) - \tilde{R}_\varphi^m(f)|. \qquad (29)
\end{aligned}$$

First two terms of (29) were upper bounded in Section B.2. Note that the upper bound of the second term (proved in Theorem 3) was based on the assumption that functionals in F_Q are Lipschitz on $\mathcal{H}_k$ w.r.t. the $\mathcal{H}_k$ metric. But as we already noted, for bounded kernels we have $\mathcal{H}_k \subseteq L_2(Q)$ which implies $\|h\|_{L_2(Q)} \leq \|h\|_{\mathcal{H}_k}$ for any $h \in \mathcal{H}_k$ (see (26)). Thus $|f(h) - f(h')| \leq L_f\|h - h'\|_{L_2(Q)} \leq L_f\|h-h'\|_{\mathcal{H}_k}$ for any $h, h' \in \mathcal{H}_k$. It means that the assumptions of Theorem 3 hold true and we can safely apply it to upper bound the first two terms of (29).

We are now going to upper bound the third one using Lemma 1:

$$\begin{aligned}
\sup_{f\in\mathcal{F}_Q}|\tilde{R}_\varphi(f) - \tilde{R}_\varphi^m(f)| &= \sup_{f\in\mathcal{F}_Q}\left|\frac{1}{n}\sum_{i=1}^n \varphi\left(-f\bigl(\mu_k(P_{S_i})\bigr)l_i\right) - \frac{1}{n}\sum_{i=1}^n\varphi\left(-f\left(\frac{1}{n_i}\sum_{z\in S_i}\hat{g}_m^z(\cdot)\right)l_i\right)\right|\\
&\leq \frac{1}{n}\sum_{i=1}^n\sup_{f\in\mathcal{F}_Q}\left|\varphi\left(-f\bigl(\mu_k(P_{S_i})\bigr)l_i\right) - \varphi\left(-f\left(\frac{1}{n_i}\sum_{z\in S_i}\hat{g}_m^z(\cdot)\right)l_i\right)\right|\\
&\leq \frac{L_\varphi}{n}\sum_{i=1}^n\sup_{f\in\mathcal{F}_Q}\left|f\bigl(\mu_k(P_{S_i})\bigr) - f\left(\frac{1}{n_i}\sum_{z\in S_i}\hat{g}_m^z(\cdot)\right)\right|\\
&\leq \frac{L_\varphi}{n}\sum_{i=1}^n\sup_{f\in\mathcal{F}_Q}L_f\left\|\mu_k(P_{S_i}) - \frac{1}{n_i}\sum_{z\in S_i}\hat{g}_m^z(\cdot)\right\|_{L_2(Q)}.
\end{aligned}$$

We can now use Lemma 1 combined in union bound over $i = 1, \ldots, n$ with $\delta' = \delta/n$. This will give us that

$$\sup_{f\in\mathcal{F}_Q}|\tilde{R}_\varphi(f) - \tilde{R}_\varphi^m(f)| \leq \frac{L_\varphi L_Q}{n}\sum_{i=1}^n\frac{2C_k}{\sqrt{m}}\left(1 + \sqrt{2\log(n\cdot n_i/\delta)}\right).$$

with probability not less than $1-\delta$ over $\{(w_i,b_i)\}_{i=1}^m$. □

C. Training and Test Protocols for Section 5.4

The synthesis of the training data for the experiments described in Section 5.4 follows a very similar procedure to the one from Section 5.1. The main difference here is that, when trying to infer the cause-effect relationship between two variables X_i and X_j belonging to a larger set of variables $X = (X_1, \ldots, X_d)$, we will have to account for the effects of possible confounders $X_k \subseteq X \setminus \{X_i, X_j\}$. For the sake of simplicity, we will only consider one-dimensional confounding effects, that is, scalar X_k.

C.1. Training Phase

To generate cause-effect pairs exhibiting every possible scalar confounding effect, we will generate data from the eight possible directed acyclic graphs depicted in Figure 4.

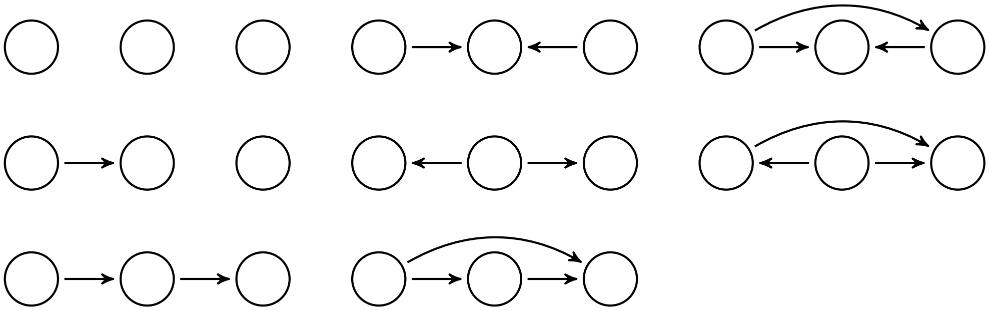

Figure 4. The eight possible directed acyclic graphs on three variables.

In particular, we will sample N different causal DAGs $G_1, \ldots, G_N$, where the G_i describes the causal structure underlying (X_i, Y_i, Z_i). Given G_i, we generate the sample set $S_i = \{(x_{ij}, y_{ij}, z_{ij})\}_{j=1}^n$ according to the generative process described in Section 5.1. Together with S_i, we annotate the triplet of labels (l_{i1}, l_{i2}, l_{i3}), where according to G_i,

- $l_{i1} = +1$ if "$X_i \to Y_i$", $l_{i1} = -1$ if "$X_i \leftarrow Y_i$", and $l_{i1} = 0$ else.
- $l_{i2} = +1$ if "$Y_i \to Z_i$", $l_{i2} = -1$ if "$Y_i \leftarrow Z_i$", and $l_{i2} = 0$ else.
- $l_{i3} = +1$ if "$X_i \to Z_i$", $l_{i1} = -1$ if "$X_i \leftarrow Z_i$", and $l_{i1} = 0$ else.

Then, we add the following six elements to our training set:

$$(\{(x_{ij}, y_{ij}, z_{ij})\}_{j=1}^n, +l_1), (\{(y_{ij}, z_{ij}, x_{ij})\}_{j=1}^n, +l_2), (\{(x_{ij}, z_{ij}, y_{ij})\}_{j=1}^n, +l_3),$$
$$(\{(y_{ij}, x_{ij}, z_{ij})\}_{j=1}^n, -l_1), (\{(z_{ij}, y_{ij}, x_{ij})\}_{j=1}^n, -l_2), (\{(z_{ij}, x_{ij}, y_{ij})\}_{j=1}^n, -l_3),$$

for all $1 \leq i \leq N$. Therefore, our training set will consist on $6N$ sample sets and their paired labels. At this point, and given any sample $\{(u_{ij}, v_{ij}, w_{ij})\}_{j=1}^n$ from the training set, we propose to use as feature vectors the concatenation of the $m-$dimensional empirical kernel mean embeddings (14) of $\{u_{ij}\}_{j=1}^n$, $\{v_{ij}\}_{j=1}^n$, and $\{(u_{ij}, v_{ij}, w_{ij})\}_{j=1}^n$, respectively.

C.2. Test Phase

To start, given n_{te} test $d-$dimensional samples $S = \{(x_{1i}, \ldots, x_{di})\}_{i=1}^{n_{te}}$, the hyper-parameters of the kernel and training data synthesis process are transductively chosen, as described in Section 5.1.

In order to estimate the causal graph underlying the test sample set S, we compute three $d \times d$ matrices $M_\to$, $M_{\perp\!\!\!\perp}$, and $M_\leftarrow$. Each of these matrices will contain, at their coordinates i, j, the probabilities of the labels "$X_i \to X_j$", "$X_i \perp\!\!\!\perp X_j$", and "$X_i \leftarrow X_j$", respectively, when averaged over all possible scalar confounders X_k. Using these matrices, we estimate the underlying causal graph by selecting the type of each edge (forward, backward, or no edge) to be the one with maximal probability according. As a post-processing step, we prune the least-confident edges until the derived graph is acyclic.