Supplementary material for: Generalizing from Several Related Classification Tasks to a New Unlabeled Sample

Gilles Blanchard
Universität Potsdam
blanchard@math.uni-potsdam.de

Gyemin Lee, Clayton Scott
University of Michigan
{gyemin,clayscot}@umich.edu

1 Background on Kernels

The function $k : \Omega \times \Omega \to \mathbb{R}$ is called a *kernel* on Ω if the matrix $(k(x_i, x_j))_{1 \le i,j \le n}$ is positive semi-definite for all positive integers n and all $x_1, \ldots, x_n \in \Omega$. It is well-known that if k is a kernel on Ω, then there exists a Hilbert space $\tilde{\mathcal{H}}$ and $\tilde{\Phi} : \Omega \to \tilde{\mathcal{H}}$ such that $k(x, x') = \langle \tilde{\Phi}(x), \tilde{\Phi}(x') \rangle_{\tilde{\mathcal{H}}}$. While $\tilde{\mathcal{H}}$ and $\tilde{\Phi}$ are not uniquely determined by k, the Hilbert space of functions $\mathcal{H}_k = \{\langle v, \tilde{\Phi}(\cdot) \rangle_{\tilde{\mathcal{H}}} : v \in \tilde{\mathcal{H}}\}$ is uniquely determined by k, and is called the reproducing kernel Hilbert space (RKHS) of k.

One way to envision $\mathcal{H}_k$ is as follows. Define $\Phi(x) := k(\cdot, x)$, which is called the *canonical feature map* associated with k. Then $\mathcal{H}_k$ is the completion of the span of $\{\Phi(x) : x \in \Omega\}$. We also recall the *reproducing property*, which states that $\langle f, \Phi(x) \rangle = f(x)$ for all $f \in \mathcal{H}_k$.

A kernel k on a compact metric space Ω is said to be *universal* when its RKHS is dense in $\mathcal{C}(\Omega)$, the set of continuous functions on Ω, with respect to the supremum norm. Universal kernels are important for establishing universal consistency of many learning algorithms.

If k is a kernel on Ω, then

$$k^*(x, x') := \frac{k(x, x')}{\sqrt{k(x, x) k(x', x')}}$$

is the associated *normalized* kernel. If a kernel is universal, then so is its associated normalized kernel. For example, the exponential kernel $k(x, x') = \exp(\kappa \langle x, x' \rangle_{\mathbb{R}^d})$, $\kappa > 0$, can be shown to be universal on $\mathbb{R}^d$ through a Taylor series argument. Consequently, the Gaussian kernel

$$k_\sigma(x, x') := \frac{\exp(\frac{1}{\sigma^2} \langle x, x' \rangle)}{\exp(\frac{1}{2\sigma^2} \|x\|^2) \exp(\frac{1}{2\sigma^2} \|x'\|^2)}$$

is universal, being the normalized kernel associated with the exponential kernel with $\kappa = 1/\sigma^2$. See [1] for additional details and discussion.

2 Implementation

We describe an implementation of our methodology for the hinge loss, $\ell(t, y) = \max(0, 1 - yt)$. To make the presentation more concise, we will employ the extended feature representation $\tilde{X} = (\widehat{P}_X, X)$, and we will also employ a single index on these variables and on the labels. Thus the training data are $(\tilde{X}_i, Y_i)_{1 \le i \le M}$, where $M = \sum_{i=1}^N n_i$, and we seek a solution to

$$\min_{f \in \mathcal{H}_{\bar{k}}} \sum_{i=1}^M c_i \max(0, 1 - Y_i f(\widetilde{X}_i)) + \frac{1}{2} \|f\|^2 .$$

Here $c_i = \frac{1}{2\lambda N n_m}$, where m is the smallest positive integer such that $i \leq n_1 + \cdots + n_m$. By the representer theorem [1], the solution of (3) has the form

$$\widehat{f}_\lambda = \sum_{i=1}^{M} r_i \overline{k}(\tilde{X}_i, \cdot)$$

for real numbers r_i. Plugging this expression into the objective function of (3), and introducing the auxiliary variables ξ_i, we have the quadratic program

$$\begin{aligned} \min_{r,\xi} \ & \frac{1}{2} r^T \overline{K} r + \sum_{i=1}^{M} c_i \xi_i \\ \text{s.t. } & Y_i \sum_{j=1}^{M} r_j \overline{k}(\tilde{X}_i, \tilde{X}_j) \geq 1 - \xi_i, \ \forall i \\ & \xi_i \geq 0, \ \forall i, \end{aligned}$$

where $\overline{K} := (\overline{k}(\tilde{X}_i, \tilde{X}_j))_{1 \leq i,j \leq M}$. Using Lagrange multiplier theory, and provided $\overline{K}$ is positive definite, the dual quadratic program is

$$\begin{aligned} \max_{\alpha} \ & -\frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j Y_i Y_j \overline{k}(\tilde{X}_i, \tilde{X}_j) + \sum_{i=1}^{M} \alpha_i \\ \text{s.t. } & 0 \leq \alpha_i \leq c_i \ \forall i, \end{aligned}$$

and the optimal function is

$$\widehat{f}_\lambda = \sum_{i=1}^{M} \alpha_i Y_i \overline{k}(\tilde{X}_i, \cdot).$$

This is equivalent to the dual of a cost-sensitive support vector machine, without offset, where the costs are given by c_i. Therefore we can learn the weights α_i using any existing software package for SVMs that accepts example-dependent costs and a user-specified kernel matrix, and allows for no offset. Returning to the original notation, the final predictor has the form

$$\widehat{f}_\lambda(\widehat{P}_X, x) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \alpha_{ij} Y_{ij} \overline{k}((\widehat{P}_X^{(i)}, X_{ij}), (\widehat{P}_X, x))$$

where the α_{ij} are nonnegative. Like the SVM, the solution is often sparse, meaning most α_{ij} are zero.

Finally, we remark on the computation of $k_P(\widehat{P}_X, \widehat{P}'_X)$. When $\mathfrak{K}$ has the form of (7) or (8), calculation of k_P may be reduced to the calculation of $\left\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \right\rangle$. If $\widehat{P}_X$ and $\widehat{P}'_X$ are based on the samples $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_{n'}$, then

$$\begin{aligned} \left\langle \Psi(\widehat{P}_X), \Psi(\widehat{P}'_X) \right\rangle &= \left\langle \frac{1}{n} \sum_{i=1}^{n} k'_X(X_i, \cdot), \frac{1}{n'} \sum_{j=1}^{n'} k'_X(X'_j, \cdot) \right\rangle \\ &= \frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} k'_X(X_i, X'_j). \end{aligned}$$

Note that when k'_X is a (normalized) Gaussian kernel, $\Psi(\widehat{P}_X)$ is just a kernel density estimate for P_X.

3 Proof of Theorem 5.1

We control the difference between the training loss and the idealized test loss via the following decomposition:

$$\begin{aligned}
&\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\ell(f(\widetilde{X}_{ij}),Y_{ij})-\mathcal{E}(f,\infty)\right| \\
&\qquad\leq \sup_{f\in\mathcal{B}_{\overline{k}}(R)}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\ell(f(\widehat{P}_X^{(i)},X_{ij}),Y_{ij})-\ell(f(P_X^{(i)},X_{ij}),Y_{ij})\Big)\right| \\
&\qquad\quad+\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\left|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\ell(f(P_X^{(i)},X_{ij}),Y_{ij})-\mathcal{E}(f,\infty)\right| \\
&\qquad=: (I)+(II).
\end{aligned}$$

3.1 Control of term (I)

Using the assumption that the loss ℓ is L_ℓ-Lipschitz in its first coordinate, we can bound the first term as follows:

$$\begin{aligned}
(I)&\leq L_\ell\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{i=1}^{n_i}\left|f(\widehat{P}_X^{(i)},X_{ij})-f(P_X^{(i)},X_{ij})\right| \\
&\leq L_\ell\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\frac{1}{N}\sum_{i=1}^{N}\left\|f(\widehat{P}_X^{(i)},.)-f(P_X^{(i)},.)\right\|_\infty
\end{aligned}$$

We now use the following result:

Lemma 3.1. *Assume the general conditions in* **(Kernels-A)** *hold. Let P_X be an arbitrary distribution on $\mathcal{X}$ and $\widehat{P}_X$ denote an empirical distribution on $\mathcal{X}$ based on an iid sample of size n from P_X. Then with probability at least $1-\delta$ over the draw of this sample, it holds that*

$$\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\left\|f(\widehat{P}_X^{(i)},.)-f(P_X^{(i)},.)\right\|_\infty\leq 3RB_kB_{k'}L_{\mathfrak{K}}\left(\frac{\log 2\delta^{-1}}{n}\right)^{\frac{\alpha}{2}}.$$

Proof. Let $X_1,\ldots,X_n$ denote the n-sample from P_X. Let us denote Φ'_X the canonical mapping $x\mapsto k'_X(x,.)$ from $\mathcal{X}$ into $\mathcal{H}_{k'_\mathcal{X}}$. We have for all $x\in\mathcal{X}$, $\|\Phi'_X(x)\|\leq B_{k'}$, so, as a consequence of Hoeffding's inequality in a Hilbert space (see, e.g., [2]) we have with probability $1-\delta$:

$$\left\|\Psi(P_X)-\Psi(\widehat{P}_X)\right\|=\left\|\frac{1}{n}\sum_{i=1}^{n}\Phi'_X(X_i)-\mathbb{E}_{X\sim P_X}\left[\Phi'_X(X)\right]\right\|\leq 3B_{k'}\sqrt{\frac{\log 2\delta^{-1}}{n}}. \tag{1}$$

On the other hand, using the reproducing property of the kernel $\overline{k}$, we have for any $x \in \mathcal{X}$ and $f \in \mathcal{B}_{\overline{k}}(R)$:

$$\begin{aligned}
|f(\widehat{P}_X, x) - f(P_X, x)| & \\
&= \left|\left\langle \overline{k}((\widehat{P}_X, x), .) - \overline{k}((P_X, x), .), f \right\rangle\right| \\
&\leq \|f\| \left\| \overline{k}((\widehat{P}_X, x), .) - \overline{k}((P_X, x), .) \right\| \\
&\leq R k_X(x,x)^{\frac{1}{2}} \Big(\mathfrak{K}(\Psi(P_X), \Psi(P_X)) \\
&\qquad + \mathfrak{K}(\Psi(\widehat{P}_X), \Psi(\widehat{P}_X)) - 2\mathfrak{K}(\Psi(P_X), \Psi(\widehat{P}_X)) \Big)^{\frac{1}{2}} \\
&\leq R B_k \left\| \Phi_{\mathfrak{K}}(\Psi(P_X)) - \Phi_{\mathfrak{K}}(\Psi(\widehat{P}_X)) \right\| \\
&\leq R B_k L_{\mathfrak{K}} \left\| \Psi(P_X) - \Psi(\widehat{P}_X) \right\|^{\alpha},
\end{aligned}$$

where we have used the fact that for all $P \in \mathfrak{P}_{\mathcal{X}}$, $\|\Psi(P)\| \leq \int_{\mathcal{X}} \|k'_X(x, \cdot)\| \, dP_X(x) \leq B_{k'}$, so that $\Psi(P) \in \mathcal{B}_{k'_X}(B_{k'})$. Combining with (1) gives the result. □

Conditionally to the draw of $(P_X^{(i)})_{1 \leq i \leq N}$, we can now apply this lemma to each $(P_X^{(i)}, \widehat{P}_X^{(i)})$ then the union bound over $i = 1, \ldots, N$ to get that with probability $1 - \delta$ (conditionally to $(P_X^{(i)})_{1 \leq i \leq N}$, and thus also unconditionally):

$$(I) \leq 3 R B_k B_{k'} L_\ell L_{\mathfrak{K}} \left(\frac{\log \delta^{-1} + \log 2N}{n} \right)^{\frac{\alpha}{2}} .$$

3.2 Control of term (II)

First define the conditional (idealized) test error for a given test distribution P_{XY}^T as

$$\mathcal{E}(f, \infty | P_{XY}^T) := \mathbb{E}_{(X^T, Y^T) \sim P_{XY}^T} \left[\ell(f(P_X^T, X^T), Y^T) \right]. \tag{2}$$

We can further decompose (II) as

$$\begin{aligned}
(II) &\leq \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \Big(\ell(f(P_X^{(i)}, X_{ij}), Y_{ij}) - \mathcal{E}(f, \infty | P_{XY}^{(i)}) \Big) + \frac{1}{N} \sum_{i=1}^{N} \Big(\mathcal{E}(f, \infty | P_{XY}^{(i)}) - \mathcal{E}(f, \infty) \Big) \\
&=: (IIa) + (IIb).
\end{aligned}$$

We recall in what follows that the loss function ℓ is positive and bounded by the constant B_ℓ, and that the kernel $\mathfrak{K}$ is bounded by $B_{\mathfrak{K}}^2$.

Control of term (IIa). We study term (IIa) conditional to $(P_{XY}^{(i)})_{1 \leq i \leq N}$. In this case, note that for this conditional distribution, the variables $(X_{ij}, Y_{ij})_{ij}$ are now independent (but not identically distributed) variables. We can thus apply the Azuma-McDiarmid inequality [3] to the function

$$\zeta((X_{ij}, Y_{ij})_{ij}) := \sup_{f \in \mathcal{B}_{\overline{k}}(R)} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \Big(\ell(f(P_X^{(i)}, X_{ij}), Y_{ij}) - \mathcal{E}(f, \infty | P_{XY}^{(i)}) \Big).$$

We deduce that with probability $1 - \delta$ over the (conditional, then also unconditional) draw of $(X_{ij}, Y_{ij})_{ij}$, it holds

$$\left| \zeta - \mathbb{E}\left[\zeta \middle| (P_{XY}^{(i)})_{1 \leq i \leq N} \right] \right| \leq \sqrt{C_\zeta \log \delta^{-1}} \, ;$$

where

$$C_\zeta := \frac{B_\ell^2}{N^2} \sum_{i=1}^{N} \frac{1}{n_i} \, ;$$

note that when all n_is are equal to n, this simplifies to

$$C_\zeta := \frac{B_\ell^2}{Nn}\,.$$

Next, to bound $\mathbb{E}\left[\zeta\middle|(P_{XY}^{(i)})_{1\le i\le N}\right]$, we can use relatively standard Rademacher complexity analysis. Denote $(\varepsilon_{ij})_{1\le i\le N, 1\le j\le n_i}$ iid Rademacher variables (independent from everything else). We have

$$\begin{aligned}
&\mathbb{E}\left[\zeta\middle|(P_{XY}^{(i)})_{1\le i\le N}\right]\\
&= \mathbb{E}_{(X_{ij},Y_{ij})}\left[\frac{1}{N}\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\ell(f(P_X^{(i)},X_{ij}),Y_{ij}) - \mathcal{E}(f,\infty|P_{XY}^{(i)})\right)\middle|(P_{XY}^{(i)})_{1\le i\le N}\right]\\
&\le \frac{2}{N}\mathbb{E}_{(X_{ij},Y_{ij})}\mathbb{E}_{(\varepsilon_{ij})}\left[\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij}\left(\ell(f(P_X^{(i)},X_{ij}),Y_{ij})\right)\middle|(P_{XY}^{(i)})_{1\le i\le N}\right]\\
&\le \frac{2RL_\ell B_k B_{\mathfrak{K}}}{N}\sqrt{\sum_{i=1}^{N}\frac{1}{n_i}}\,.
\end{aligned}$$

The first inequality is a standard symmetrization argument. The last inequality is a variation (with possibly unequal weights $1/n_i$) on the standard bound (see [4], Theorem 7 and Lemma 22) for the Rademacher complexity of a Lipschitz loss function on the ball of radius R of $\mathcal{H}_{\overline{k}}$, the kernel $\overline{k}$ being bounded by $B_k^2 B_{\mathfrak{K}}^2$. In case all n_is are equal, this boils down to

$$\mathbb{E}\left[\zeta\middle|(P_{XY}^{(i)})_{1\le i\le N}\right] \le 2L_\ell R B_X B_{\mathfrak{K}}\sqrt{\frac{1}{Nn}}\,.$$

Control of term (IIb). Since the $(P_{XY}^{(i)})_{1\le i\le N}$ are iid, we can apply the Azuma-McDiarmid inequality to the function

$$\xi((P_{XY}^{(i)})_{1\le i\le N}) := \sup_{f\in\mathcal{B}_{\overline{k}}(R)}\frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{E}(f,\infty|P_{XY}^{(i)}) - \mathcal{E}(f,\infty)\right)\,,$$

obtaining that with probability $1-\delta$ over the draw of $(P_{XY}^{(i)})_{1\le i\le N}$, it holds

$$|\xi - \mathbb{E}[\xi]| \le B_\ell\sqrt{\frac{\log\delta^{-1}}{2N}}\,;$$

Rademacher complexity analysis for bounding $\mathbb{E}[\xi]$: below, we will denote (X_i,Y_i) a (single) draw from distribution $P_{XY}^{(i)}$ (and these draws are independent). We also denote $(\varepsilon_i)_{1\le i\le N}$ iid Rademacher variables (independent from everything else). We have

$$\begin{aligned}
\mathbb{E}[\xi] &= \mathbb{E}_{(P_{XY}^{(i)})_{1\le i\le N}}\left[\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{(X,Y)\sim P_{XY}^{(i)}}\left[\ell(f(P_X,X),Y)\right]\right.\\
&\qquad\left. - \mathbb{E}_{P_{XY}\sim\mu}\mathbb{E}_{(X,Y)\sim P_{XY}}\left[\ell(f(P_X,X),Y)\right]\right]\\
&\le \frac{2}{N}\mathbb{E}_{(P_{XY}^{(i)})_{1\le i\le N}}\mathbb{E}_{(\varepsilon_i)_{1\le i\le N}}\left[\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\sum_{i=1}^{N}\varepsilon_i\mathbb{E}_{(X_i,Y_i)\sim P_{XY}^{(i)}}\left[\ell(f(P_X^{(i)},X_i),Y_i)\right]\right]\\
&\le \frac{2}{N}\mathbb{E}_{(P_{XY}^{(i)})_{1\le i\le N}}\mathbb{E}_{(X_i,Y_i)_{1\le i\le N}}\mathbb{E}_{(\varepsilon_i)_{1\le i\le N}}\left[\sup_{f\in\mathcal{B}_{\overline{k}}(R)}\sum_{i=1}^{N}\varepsilon_i\ell(f(P_X^{(i)},X_i),Y_i)\right]\\
&\le \frac{2RL_\ell B_k B_{\mathfrak{K}}}{\sqrt{N}}\,.
\end{aligned}$$

The first inequality is a standard symmetrization argument. In the second inequality, the inner expectation on the (X_i, Y_i) is pulled outwards. The last inequality is a standard bound for the Rademacher complexity of a Lipschitz loss function on the ball of radius R of $\mathcal{H}_{\overline{k}}$, the kernel $\overline{k}$ being bounded by $B_k^2 B_{\mathfrak{K}}^2$.

Combining all of the above inequalities, we obtained the announced result of the theorem.

3.3 Regularity conditions for the kernel $\mathfrak{K}$

We investigate sufficient conditions on the kernel $\mathfrak{K}$ to ensure the regularity condition (9). Roughly speaking, the regularity of the feature mapping of a reproducing kernel is "one half" of the regularity of the kernel in each of its variables. The next result considers the situation where $\mathfrak{K}$ is itself simply a Hölder continuous function of its variables.

Lemma 3.2. *Let $\alpha \in (0, \frac{1}{2}]$. Assume that the kernel $\mathfrak{K}$ is Hölder continuous of order 2α and constant $L_{\mathfrak{K}}^2/2$ in each of its two variables on $\mathcal{B}_{k'_X}(B_{k'})$. Then (9) is satisfied.*

Proof. For any $v, w \in \mathcal{B}_{k'_X}(B_{k'})$:

$$\|\Phi_{\mathfrak{K}}(v) - \Phi_{\mathfrak{K}}(w)\| = \left(\mathfrak{K}(v,v) + \mathfrak{K}(w,w) - 2\mathfrak{K}(v,w)\right)^{\frac{1}{2}} \leq L_{\mathfrak{K}} \|v - w\|^{\frac{\alpha}{2}}$$

□

The above type of regularity only leads to a Hölder feature mapping of order at most $\frac{1}{2}$ (when the kernel function is Lipschitz continuous in each variable). Since this order plays an important role in the rate of convergence of the upper bound in the main error control theorem, it is desirable to study conditions ensuring more regularity, in particular a feature mapping which has at least Lipschitz continuity. For this, we consider the following stronger condition, namely that the kernel function is twice differentiable in a specific sense:

Lemma 3.3. *Assume that, for any $u, v \in \mathcal{B}_{k'_X}(B_{k'})$ and unit norm vector e of $\mathcal{H}_{k'_X}$, the function $h_{u,v,e} : (\lambda, \mu) \in \mathbb{R}^2 \mapsto \mathfrak{K}(u + \lambda e, v + \mu e)$ admits a mixed partial derivative $\partial_1 \partial_2 h_{u,v,e}$ at the point $(\lambda, \mu) = (0, 0)$ which is bounded in absolute value by a constant $C_{\mathfrak{K}}^2$ independently of (u, v, e).*

Then (9) is satisfied with $\alpha = 1$ and $L_{\mathfrak{K}} = C_{\mathfrak{K}}$, that is, the canonical feature mapping of $\mathfrak{K}$ is Lipschitz continuous on $\mathcal{B}_{k'_X}(B_{k'})$.

Proof. The argument is along the same lines as [1], Lemma 4.34. Observe that, since $h_{u,v,e}(\lambda + \lambda', \mu + \mu') = h_{u+\lambda e, v+\mu e, e}(\lambda', \mu')$, the function $h_{u,v,e}$ actually admits a uniformly bounded mixed partial derivative in any point $(\lambda, \mu) \in \mathbb{R}^2$ such that $(u + \lambda e, v + \mu e) \in \mathcal{B}_{k'_X}(B_{k'})$. Let us denote $\Delta_1 h_{u,v,e}(\lambda, \mu) := h_{u,v,e}(\lambda, \mu) - h_{u,v,e}(0, \mu)$. For any $u, v \in \mathcal{B}_{k'_X}(B_{k'})$, $u \neq v$, let us denote $\lambda := \|v - u\|$ and the unit vector $e := \lambda^{-1}(v - u)$; we have

$$\begin{aligned}
\|\Phi_{\mathfrak{K}}(u) - \Phi_{\mathfrak{K}}(v)\|^2 &= \mathfrak{K}(u,u) + \mathfrak{K}(u + \lambda e, u + \lambda e) - \mathfrak{K}(u, u + \lambda e) - \mathfrak{K}(u + \lambda e, u) \\
&= \Delta_1 h_{u,v,e}(\lambda, \lambda) - \Delta_1 h_{u,v,e}(\lambda, 0) \\
&= \lambda \partial_2 \Delta_1 h_{u,v,e}(\lambda, \lambda'),
\end{aligned}$$

where we have used the mean value theorem, yielding existence of $\lambda' \in [0, \lambda]$ such that the last equality holds. Furthermore,

$$\begin{aligned}
\partial_2 \Delta_1 h_{u,v,e}(\lambda, \lambda') &= \partial_2 h_{u,v,e}(\lambda, \lambda') - \partial_2 h_{u,v,e}(0, \lambda') \\
&= \lambda \partial_1 \partial_2 h_{u,v,e}(\lambda'', \lambda'),
\end{aligned}$$

using again the mean value theorem, yielding existence of $\lambda'' \in [0, \lambda]$ in the last equality. Finally, we get

$$\|\Phi_{\mathfrak{K}}(u) - \Phi_{\mathfrak{K}}(v)\|^2 = \lambda^2 \partial_1 \partial_2 h_{u,v,e}(\lambda', \lambda'') \leq C_{\mathfrak{K}}^2 \|v - u\|^2 .$$

□

Lemma 3.4. *Assume that the kernel $\mathfrak{K}$ takes the form of either (a) $\mathfrak{K}(u,v) = g(\|u-v\|^2)$ or (b) $\mathfrak{K}(u,v) = g(\langle u,v\rangle)$, where g is a twice differentiable real function of real variable defined on $[0, 4B_{k'}^2]$ in case (a), and on $[-B_{k'}^2, B_{k'}^2]$ in case (b). Assume $\|g'\|_\infty \leq C_1$ and $\|g''\|_\infty \leq C_2$. Then $\mathfrak{K}$ satisfies the assumption of Lemma 3.3 with $C_{\mathfrak{K}} := 2C_1 + 16C_2B_{k'}^2$ in case (a), and $C_{\mathfrak{K}} := C_1 + B_{k'}^2 C_2$ for case (b).*

Proof. In case (a), we have $h_{u,v,e}(\lambda,\mu) = g(\|u-v+(\lambda-\mu)e\|^2)$. It follows

$$\begin{aligned}|\partial_1\partial_2 h_{u,v,e}(0,0)| &= \left|-2g'(\|u-v\|^2)\,\|e\|^2 - 4g''(\|u-v\|^2)\,\langle u-v,e\rangle^2\right| \\ &\leq 2C_1 + 16B_{k'}^2C_2\,.\end{aligned}$$

In case (b), we have $h_{u,v,e}(\lambda,\mu) = g(\langle u+\lambda e, v+\mu e\rangle)$. It follows

$$\begin{aligned}|\partial_1\partial_2 h_{u,v,e}(0,0)| &= \left|g'(\langle u,v\rangle)\,\|e\|^2 + g''(\langle u,v\rangle)\,\langle u,e\rangle\,\langle v,e\rangle\right| \\ &\leq C_1 + B_{k'}^2C_2\,.\end{aligned}$$

□

3.4 Proof of Lemma 5.2

Proof. Let $\mathcal{H}, \mathcal{H}'$ the RKHS associated to k, k' with the associated feature mappings Φ, Φ'. Then it can be checked that $(x,x') \in \mathcal{X}\times\mathcal{X}' \mapsto \Phi(x)\otimes\Phi'(x')$ is a feature mapping for $\overline{k}$ into the Hilbert space $\mathcal{H}\otimes\mathcal{H}'$. Using [1], Th. 4.21, we deduce that the RKHS $\overline{H}$ of $\overline{k}$ contains precisely all functions of the form $(x,x') \in \mathcal{X}\times\mathcal{X}' \mapsto F_w(x,x') = \langle w, \Phi(x)\otimes\Phi(x')\rangle$, where w ranges over $\mathcal{H}\otimes\mathcal{H}'$. Taking w of the form $w = g\otimes g'$, $g\in\mathcal{H}$, $g\in\mathcal{H}'$, we deduce that $\overline{H}$ contains in particular all functions of the form $f(x,x') = g(x)g(x')$, and further

$$\widetilde{\mathcal{H}} := \operatorname{span}\{(x,x') \in \mathcal{X}\times\mathcal{X}' \mapsto g(x)g(x'); g\in\mathcal{H}, g'\in\mathcal{H}'\} \subset \overline{H}.$$

Denote $\mathcal{C}(\mathcal{X}), \mathcal{C}(\mathcal{X}'), \mathcal{C}(\mathcal{X}\times\mathcal{X}')$ the set of real-valued continuous functions on the respective spaces. Let

$$\mathcal{C}(\mathcal{X})\otimes\mathcal{C}(\mathcal{X}') := \operatorname{span}\{(x,x') \in \mathcal{X}\times\mathcal{X}' \mapsto f(x)f'(x'); f\in\mathcal{C}(\mathcal{X}), f'\in\mathcal{C}(\mathcal{X}')\}\,.$$

Let $G(x,x')$ be an arbitrary element of $\mathcal{C}(\mathcal{X})\otimes\mathcal{C}(\mathcal{X}')$, $G(x,x') = \sum_{i=1}^k \lambda_i g_i(x)g_i'(x')$ with $g_i \in \mathcal{C}(\mathcal{X}), g_i' \in \mathcal{C}(\mathcal{X}')$ for $i = 1,\dots,k$. For $\varepsilon > 0$, by universality of k and k', there exist $f_i \in \mathcal{H}, f_i' \in \mathcal{H}'$ so that $\|f_i - g_i\|_\infty \leq \varepsilon$, $\|f_i' - g_i'\|_\infty \leq \varepsilon$ for $i = 1,\dots,k$. Let $F(x,x') := \sum_{i=1}^k \lambda_i f_i(x)f_i'(x') \in \widetilde{\mathcal{H}}$. We have

$$\begin{aligned}\|F(x,x') - G(x,x')\|_\infty &\leq \left\|\sum_{i=1}^k \lambda_i(g_i(x)g_i'(x) - f_i(x)f_i'(x))\right\|_\infty \\ &= \Bigg\|\sum_{i=1}^k \lambda_i\Big[(f_i(x)-g_i(x))(g_i'(x')-f_i'(x')) \\ &\qquad + g_i(x)(g_i'(x)-f_i'(x')) + (g_i(x)-f_i(x))g_i'(x')\Big]\Bigg\|_\infty \\ &\leq \varepsilon\sum_{i=1}^k |\lambda_i|\,(\varepsilon + \|g_i\|_\infty + \|g_i'\|_\infty)\,.\end{aligned}$$

This establishes that $\widetilde{\mathcal{H}}$ is dense in $\mathcal{C}(\mathcal{X})\otimes\mathcal{C}(\mathcal{X}')$ for the supremum norm. It can be easily checked that $\mathcal{C}(\mathcal{X})\otimes\mathcal{C}(\mathcal{X}')$ is an algebra of functions which does not vanish and separates points on $\mathcal{X}\times\mathcal{X}'$. By the Stone-Weiertrass theorem, it is therefore dense in $\mathcal{C}(\mathcal{X}\times\mathcal{X}')$ for the supremum norm. We deduce that $\widetilde{\mathcal{H}}$ (and thus also $\overline{\mathcal{H}}$) is dense in $\mathcal{C}(\mathcal{X}\times\mathcal{X}')$, so that $\overline{k}$ is universal. □

3.5 Proof of Theorem 5.3

By Lemma 5.2, it suffices to show $\mathfrak{P}_{\mathcal{X}}$ is a compact metric space, and that $k_P(P_X, P'_X)$ is universal on $\mathfrak{P}_{\mathcal{X}}$. The former statement follows from Theorem 6.4 of [5], where the metric is the Prohorov metric. We will deduce the latter statement from of Theorem 2.2 of [6]. The statement of Theorem 2.2 there is in principle restricted to kernels of the form (8), but the proof actually only uses that the kernel $\mathfrak{K}$ is universal on any compact set of $\mathcal{H}_{k'_X}$. To apply Theorem 2.2, it remains to show that $\mathcal{H}_{k'_X}$ is a separable Hilbert space, and that Ψ is injective and continuous. Injectivity of Ψ is equivalent to k'_X being a characteristic kernel, which follows from the assumed universality of k'_X [7]. The continuity of k'_X implies separability of $\mathcal{H}_{k'_X}$ ([1], Lemma 4.33) as well as continuity of $\tilde{\Psi}$ ([6], Lemma 2.3 and preceding discussion). Now Theorem 2.2 of [6] may be applied, and the result follows.

The fact that kernels of the form (9), where G is analytic with positive Taylor coefficients, are universal on any compact set of $\mathcal{H}_{k'_X}$ was established in the proof of Theorem 2.2 of the same work [6].

3.6 Proof of Corollary 5.4

Proof. It has been established that $\overline{k}$ is a universal kernel on $\mathfrak{P}_X \times \mathcal{X}$. The rest of the argument is standard and we only sketch it here for completeness. Let us denote $\widehat{\mathcal{E}}(f, N, n)$ the empirical loss of a function f appearing in the LHS of (8) and in the optimization problem (3) defining the estimator. By comparing the objective function in (3) taken at the minimizer $\widehat{f}_\lambda$ and at the null function, we deduce that we must have $\left\|\widehat{f}_\lambda\right\| \leq B_\ell/\lambda$. Denote $j = \min(N, n^\alpha)$. We apply the result of Theorem 5.1 for $R_j = B_\ell/\lambda_j, \delta_j = 1/j^2$. The assumptions on λ_j ensure that the RHS of the main bound goes to 0 as $j \to \infty$. Since $\widehat{f}_{\lambda_j} \in \mathcal{B}_{\overline{k}}(R_j)$, we deduce that $\left|\widehat{\mathcal{E}}(\widehat{f}_{\lambda_j}, N, n) - \mathcal{E}(\widehat{f}_{\lambda_j}, \infty)\right|$ converges to zero in probability as $j = \min(N, n^\alpha) \to \infty$. Similarly, for any arbitrary function $f_0 \in \mathcal{H}_{\overline{k}}$, $\left|\widehat{\mathcal{E}}(f_0, N, n) - \mathcal{E}(f_0, \infty)\right|$ converges to zero in probability since $f_0 \in \mathcal{B}_{\overline{k}}(R_j)$ for j big enough (since $R_j \to \infty$ as $j \to \infty$). By comparing the objective in $\widehat{f}_{\lambda_j}$ and in f_0, and using the above bounds, we deduce

$$\mathcal{E}(\widehat{f}_{\lambda_j}, \infty) \leq \mathcal{E}(f_0, \infty) + \lambda_j \|f_0\| + \varepsilon_j,$$

where ε_j tends to 0 in probability. Since this is valid for any $f_0 \in \mathcal{H}_k$, we get

$$\mathcal{E}(\widehat{f}_{\lambda_{\min(N,n^\alpha)}}, \infty) \xrightarrow{P} \inf_{f_0 \in \mathcal{H}_k} \mathcal{E}(f_0, \infty) = \inf_{f : \mathfrak{P}_X \times \mathcal{X} \to \mathbb{R}} \mathcal{E}(f, \infty),$$

where the last equality is a consequence of the universality of $\overline{k}$ and the boundedness of ℓ. □

References

[1] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, 2008.

[2] I.F. Pinelis and A.I. Sakhanenko, "Remarks on inequalities for probabilities of large deviations," *Theory Probab. Appl.*, vol. 30, no. 1, pp. 143–148, 1985.

[3] C. McDiarmid, "On the method of bounded differences," *Surveys in Combinatorics*, vol. 141, pp. 148–188, 1989.

[4] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[5] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, 1967.

[6] A. Christmann and I. Steinwart, "Universal kernels on non-standard input spaces," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 406–414.

[7] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.