Appendix

The appendix has two major parts: proof for all the theorems and more detailed experiments (Appendix E).

A. Proofs

Proposition 1. $R(f)$ satisfies Assumption 1 if, and only if, $R(f) = \sup_{g\in S} \langle f, g\rangle_{\mathcal{H}}$, where $S \subseteq \mathcal{H}$ is bounded in the RKHS norm and is symmetric ($g \in S \Leftrightarrow -g \in S$).

Recall

Assumption 1. We assume that $R : \mathcal{F} \to \mathbb{R}$ is a semi-norm. Equivalently, $R : \mathcal{F} \to \mathbb{R}$ is convex and $R(\alpha f) = |\alpha|\, R(f)$ for all $f \in \mathcal{F}$ and $\alpha \in \mathbb{R}$ (absolute homogeneity). Furthermore, we assume R is closed (i.e., lower semicontinuous) *w.r.t.* the topology in $\mathcal{H}$.

Proposition 1 (in a much more general form), to our best knowledge, is due to Hörmander (1954). We give a "modern" proof below for the sake of completeness.

Proof for Proposition 1.
The "if" part: convexity and absolute homogeneity are trivial. To show the lower semicontinuity, we just need to show the epigraph is closed. Let (f_n, t_n) be a convergent sequence in the epigraph of R, and the limit is (f, t). Then $\langle f_n, g\rangle_{\mathcal{H}} \le t_n$ for all n and $g \in S$. Tending n to infinty, we get $\langle f, g\rangle_{\mathcal{H}} \le t$. Take supremum over g on the left-hand side, and we obtain $R(f) \le t$, i.e., (f, t) is in the epigraph of R.

The "only if" part: A sublinear function R vanishing at the origin is a support function if, and only if, it is closed. Indeed, if R is closed, then its conjugate function

$$\lambda R^*(f^*) = \lambda \left(\sup_f \langle f, f^*\rangle_{\mathcal{H}} - R(f) \right) \tag{38}$$

$$= \sup_f \langle \lambda f, f^*\rangle_{\mathcal{H}} - R(\lambda f) \tag{39}$$

$$= R^*(f^*), \tag{40}$$

is scaling invariant for any positive λ, i.e., R^* is an indicator function. Conjugating again we have $R = (R^*)^*$ is a support function. So, R is the support function of

$$S = \operatorname{dom}(R^*) = \{g : \langle f, g\rangle_{\mathcal{H}} \le R(f) \text{ for all } f \in \mathcal{H}\},$$

which is obviously closed. S is also symmetric, because the symmetry of R implies the same for its conjugate function R^*, hence its domain S.

To see S is bounded, assume to the contrary we have $\lambda_n g_n \in S$ with $\|g_n\|_{\mathcal{H}} = 1$ and $\lambda_n \to \infty$. Since R is finite-valued and closed, it is continuous, see (e.g. Borwein and Vanderwerff, 2010, Proposition 4.1.5). Thus, for any $\delta > 0$ there exists some $\epsilon > 0$ such that $\|f\|_{\mathcal{H}} \le \epsilon \implies R(f) \le \delta$. Choose $f = \epsilon g_n$ in the definition of S above we have:

$$\epsilon \lambda_n = \langle \epsilon g_n, \lambda_n g_n\rangle_{\mathcal{H}} \le R(\epsilon g_n) \le \delta, \tag{41}$$

which is impossible as $\lambda_n \to \infty$. □

Proof of Theorem 4.
a): since $\sum_i \alpha_i G^*_{x_i} = \sum_j \beta_j G^*_{z_j}$, it holds that

$$\left\langle h; \sum_i \alpha_i G^*_{x_i} \right\rangle = \left\langle h; \sum_j \beta_j G^*_{z_j} \right\rangle, \ \forall\, h \in \mathcal{F} \tag{42}$$

which implies that

$$\sum_i \alpha_i h(x_i) = \sum_j \beta_i h(z_j), \quad \forall\, h \in \mathcal{F}. \tag{43}$$

Therefore

$$\sum_i \alpha_i k(x_i, \cdot) = \sum_j \beta_j k(z_j, \cdot). \tag{44}$$

Then apply the linear map T on both sides, and we immediately get $\sum_i \alpha_i \tilde{k}_{x_i} = \sum_j \beta_j \tilde{k}_{z_j}$.

b): suppose otherwise that the completion of $\operatorname{span}\{G^*_x : x \in \mathcal{X}\}$ is not $\mathcal{B}^*$. Then by the Hahn-Banach theorem, there exists a nonzero function $f \in \mathcal{B}$ such that $\langle f; G^*_x\rangle = 0$ for all $x \in \mathcal{X}$. By (8), this means $f(x) = 0$ for all x. Since $\mathcal{B}$ is a Banach space of functions on $\mathcal{X}$, $f = 0$ in $\mathcal{B}$. Contradiction.

The linearity of ι^* follows directly from a) and b). □

To prove Theorem 5, we first introduce five lemmas. To start with, we set up the concept of *polar operator* that will be used extensively in the proof:

$$\mathrm{PO}_{\tilde{\mathsf{B}}}(u) := \arg\max_{v \in \tilde{\mathsf{B}}} \langle v, u\rangle, \quad \forall\, u \in \mathbb{R}^d. \tag{45}$$

Here the optimization is convex, and the argmax is uniquely attained because $\tilde{\mathsf{B}}$ is strictly convex. So $\|\cdot\|_{\tilde{\mathcal{B}}^*}$ is differentiable at all u, and the gradient is

$$\nabla \|u\|_{\tilde{\mathcal{B}}^*} = \mathrm{PO}_{\tilde{\mathsf{B}}}(u). \tag{46}$$

Lemma 1. *Under Assumptions 2 and 3,*

$$\|g\|_{\mathcal{B}} = \|g^*\|_{\mathcal{B}^*} = \|\iota^*(g^*)\|_{\tilde{\mathcal{B}}^*} = \|\iota(g)\|_{\tilde{\mathcal{B}}}, \quad \forall\, g \in \mathcal{B}. \tag{47}$$

Proof. The first equality is trivial, and the third equality is by the definition of $\iota(g)$ in (23). To prove the second equality, let us start by considering $g^* = \sum_i \alpha_i G^*_{x_i}$. Then

$$\|\iota^*(g^*)\|_{\tilde{\mathcal{B}}^*} = \max_{v\in\tilde{\mathsf{B}}} \langle v, \iota^*(g^*)\rangle \quad (48)$$

$$= \max_{v\in\tilde{\mathsf{B}}} \sum_i \alpha_i \left\langle v, \tilde{k}_{x_i}\right\rangle \quad (49)$$

$$\|g^*\|_{\mathcal{B}^*} = \max_{f\in\mathsf{B}} \langle f; g^*\rangle = \max_{f\in\mathsf{B}} \sum_i \alpha_i f(x_i) \quad (50)$$

$$= \max_{f\in\mathsf{B}} \sum_i \alpha_i \langle f, k(x_i,\cdot)\rangle_{\mathcal{H}} \quad (51)$$

$$= \max_{f\in\mathsf{B}} \sum_i \alpha_i \left\langle \tilde{f}, \tilde{k}_{x_i}\right\rangle, \quad (52)$$

where the last equality is by Assumption 3. So it suffices to show that $\tilde{\mathsf{B}} = \{\tilde{f} : f\in\mathsf{B}\}$.

"$\supseteq$" is trivial because for all $f\in\mathsf{B}$, by Assumption 3,

$$\left\|\tilde{f}\right\|^2 + \max_{z\in S}\left\langle \tilde{z}, \tilde{f}\right\rangle^2 = \|f\|^2_{\mathcal{H}} + \max_{z\in S}\langle z, f\rangle^2_{\mathcal{H}} \le 1. \quad (53)$$

"$\subseteq$": for any $v\in\tilde{\mathsf{B}}$, Assumption 2 asserts that there exists $h_v\in\mathcal{H}$ such that $\tilde{h}_v = v$. Then by Assumption 3,

$$\|h_v\|^2_{\mathcal{H}} + \max_{z\in S}\langle z, h_v\rangle^2_{\mathcal{H}} = \|v\|^2 + \max_{z\in S}\langle \tilde{z}, v\rangle^2 \le 1. \quad (54)$$

Since both $\|\cdot\|_{\mathcal{B}^*}$ and $\|\cdot\|_{\tilde{\mathcal{B}}^*}$ are continuous, applying the denseness result in part b) of Theorem 4 completes the proof of the second equality in (47). □

Lemma 2. *Under Assumptions 2 and 3,*

$$\langle \iota(f), \iota^*(g^*)\rangle = \langle f; g^*\rangle, \quad \forall\, f\in\mathcal{B}, g^*\in\mathcal{B}^*. \quad (55)$$

Proof.

$$\langle f; g^*\rangle \overset{\text{by (7)}}{=} [g^*, f^*]_{\mathcal{B}^*} \quad (56)$$

$$= \lim_{t\to 0}\frac{1}{2t}\left(\|f^* + tg^*\|^2_{\mathcal{B}^*} - \|f^*\|^2_{\mathcal{B}^*}\right) \text{ (by Giles (1967))} \quad (57)$$

$$= \lim_{t\to 0}\frac{1}{2t}\left[\|\iota^*(f^*) + t\iota^*(g^*)\|^2_{\tilde{\mathcal{B}}^*} - \|\iota^*(f^*)\|^2_{\tilde{\mathcal{B}}^*}\right], \quad (58)$$

where the last equality is by Lemma 1 and Theorem 4. Now it follows from the polar operator as discussed above that

$$\langle f; g^*\rangle = \left\langle \|\iota^*(f^*)\|_{\tilde{\mathcal{B}}^*}\cdot \mathrm{PO}_{\tilde{\mathsf{B}}}(\iota^*(f^*)), \iota^*(g^*)\right\rangle \quad (59)$$
$$= \langle \iota(f), \iota^*(g^*)\rangle.$$
□

Lemma 3. *Under Assumptions 2 and 3,*

$$\tilde{\mathsf{B}} = \iota(\mathsf{B}) := \{\iota(f) : \|f\|_{\mathcal{B}} \le 1\}. \quad (60)$$

Proof. "LHS $\supseteq$ RHS": by Lemma 1, it is obvious that $\|f\|_{\mathcal{B}} \le 1$ implies $\|\iota(f)\|_{\tilde{\mathcal{B}}} \le 1$.

"LHS $\subseteq$ RHS": we are to show that for all $v\in\tilde{\mathsf{B}}$, there must exist a $f_v\in\mathsf{B}$ such that $v = \iota(f)$. If $v = 0$, then trivially set $f_v = 0$. In general, due to the polar operator definition (45), there must exist $u\in\mathbb{R}^d$ such that

$$v/\|v\|_{\tilde{\mathcal{B}}} = \mathrm{PO}_{\tilde{\mathsf{B}}}(u). \quad (61)$$

We next reverse engineer a $q^*\in\mathcal{B}^*$ so that $\iota^*(g^*) = u$. By Assumption 2, there exists $h_u\in\mathcal{H}$ such that $\tilde{h}_u = u$. Suppose $h_u = \sum_i \alpha_i k_{x_i}$. Then define $q^* = \sum_i \alpha_i G^*_{x_i}$, and we recover u by

$$\iota^*(q^*) = \sum_i \alpha_i \tilde{k}_i = \tilde{h}_u = u. \quad (62)$$

Apply Lemma 1 and we obtain

$$\|q\|_{\mathcal{B}} = \|\iota^*(q^*)\|_{\tilde{\mathcal{B}}^*} = \|u\|_{\tilde{\mathcal{B}}^*}. \quad (63)$$

Now construct

$$f_v = \frac{\|v\|_{\tilde{\mathcal{B}}}}{\|q\|_{\mathcal{B}}} q. \quad (64)$$

We now verify that $v = \iota(f_v)$. By linearity of ι^*,

$$\iota^*(f_v^*) = \frac{\|v\|_{\tilde{\mathcal{B}}}}{\|q\|_{\mathcal{B}}}\iota^*(q^*) = \frac{\|v\|_{\tilde{\mathcal{B}}}}{\|q\|_{\mathcal{B}}} u. \quad (65)$$

So $\mathrm{PO}_{\tilde{\mathsf{B}}}(\iota^*(f_v^*)) = v/\|v\|_{\tilde{\mathcal{B}}}$ and plugging into (23),

$$\iota(f_v) = \|\iota^*(f_v^*)\|_{\tilde{\mathcal{B}}^*}\,\mathrm{PO}_{\tilde{\mathsf{B}}}(\iota^*(f_v^*)) \quad (66)$$

$$= \frac{\|v\|_{\tilde{\mathcal{B}}}}{\|q\|_{\mathcal{B}}}\|u\|_{\tilde{\mathcal{B}}^*}\frac{1}{\|v\|_{\tilde{\mathcal{B}}}} v \quad (67)$$

$$= v. \quad \text{(by (63))}$$
□

Lemma 4. *Under Assumptions 2 and 3,*

$$\tilde{\mathsf{B}}^* = \iota^*(\mathsf{B}^*) := \{\iota^*(g^*) : \|g^*\|_{\mathcal{B}^*} \le 1\}. \quad (68)$$

Proof. "LHS $\supseteq$ RHS": By definition of dual norm, any $g^*\in\mathsf{B}^*$ must satisfy

$$\langle f; g^*\rangle \le 1, \quad \forall\, f\in\mathsf{B}. \quad (69)$$

Again, by the definition of dual norm, we obtain

$$\|\iota^*(g^*)\|_{\tilde{\mathcal{B}}^*} = \sup_{v\in\tilde{\mathsf{B}}}\langle v, \iota^*(g^*)\rangle \quad (70)$$

$$= \sup_{f\in\mathsf{B}}\langle \iota(f), \iota^*(g^*)\rangle \quad \text{(Lemma 3)} \quad (71)$$

$$= \sup_{f\in\mathsf{B}}\langle f; g^*\rangle \quad \text{(by Lemma 2)} \quad (72)$$

$$\le 1. \quad (73)$$

"LHS $\subseteq$ RHS": Any $u\in\mathbb{R}^d$ with $\|u\|_{\tilde{\mathcal{B}}^*} = 1$ must satisfy

$$\max_{v\in\tilde{\mathsf{B}}}\langle u, v\rangle = 1. \quad (74)$$

Denote $v = \arg\max_{v \in \tilde{\mathsf{B}}} \langle u, v \rangle$ which must be uniquely attained. So $\|v\|_{\tilde{\mathcal{B}}} = 1$. Then Lemma 3 implies that there exists a $f \in \mathsf{B}$ such that $\iota(f) = v$. By duality,

$$\max_{u \in \tilde{\mathsf{B}}^*} \langle v, u \rangle = 1, \tag{75}$$

and u is the unique maximizer. Now note

$$\langle v, \iota^*(f^*) \rangle = \langle \iota(f), \iota^*(f^*) \rangle = \langle f; f^* \rangle = 1, \tag{76}$$

where the last equality is derived from Lemma 1 with

$$\|f\|_{\mathcal{B}} = \|\iota(f)\|_{\tilde{\mathcal{B}}} = \|v\|_{\tilde{\mathcal{B}}} = 1. \tag{77}$$

Note from Lemma 1 that $\|\iota^*(f^*)\|_{\tilde{\mathcal{B}}^*} = \|f\|_{\mathcal{B}} = 1$. So $\iota^*(f^*)$ is a maximizer in (75), and as a result, $u = \iota^*(f^*)$.

If $\|u\|_{\tilde{\mathcal{B}}^*} < 1$, then just construct f as above for $u / \|u\|_{\tilde{\mathcal{B}}^*}$, and then multiply it by $\|u\|_{\tilde{\mathcal{B}}^*}$. The result will meet our need thanks to the linearity of ι^* from Theorem 4. □

Lemma 5. *Under Assumptions 2 and 3,*

$$\max_{v \in \tilde{\mathsf{B}}} \langle v, \iota^*(g^*) \rangle = \max_{f \in \mathsf{B}} \langle f; g^* \rangle, \quad \forall g^* \in \mathcal{B}^*. \tag{78}$$

Moreover, by Theorem 3, the argmax of the RHS is uniquely attained at $f = g / \|g\|_{\mathcal{B}}$, and the argmax of the LHS is uniquely attained at $v = \iota(g) / \|\iota(g)\|_{\tilde{\mathcal{B}}}$.

Proof. LHS $\geq$ RHS: Let f^{opt} be an optimal solution to the RHS. Then by Lemma 3, $\iota(f^{opt}) \in \tilde{\mathsf{B}}$, and so

$$\begin{aligned}
\text{RHS} &= \left\langle f^{opt}; g^* \right\rangle \qquad\qquad\qquad\qquad\quad \tag{79}\\
&= \left\langle \iota(f^{opt}), \iota^*(g^*) \right\rangle \quad \text{(by Lemma 2)} \tag{80}\\
&\leq \max_{v \in \tilde{\mathsf{B}}} \langle v, \iota^*(g^*) \rangle \tag{81}\\
&= \text{LHS}. \tag{82}
\end{aligned}$$

LHS $\leq$ RHS: let v^{opt} be an optimal solution to the LHS. Then by Lemma 3, there is $f_{v^{opt}} \in \mathsf{B}$ such that $\iota(f_{v^{opt}}) = v^{opt}$. So

$$\begin{aligned}
\text{LHS} &= \left\langle v^{opt}, \iota^*(g^*) \right\rangle \tag{83}\\
&= \langle \iota(f_{v^{opt}}), \iota^*(g^*) \rangle \tag{84}\\
&= \langle f_{v^{opt}}; g^* \rangle \quad \text{(by Lemma 2)} \tag{85}\\
&\leq \max_{f \in \mathsf{B}} \langle f; g^* \rangle \quad (\text{since } f_{v^{opt}} \in \mathsf{B}) \tag{86}\\
&= \text{RHS}.
\end{aligned}$$
□

Proof of Theorem 5. Let $f \in \mathcal{B}$ and $\alpha \in \mathbb{R}$. Then $(\alpha f)^* = \alpha f^*$, and by (23) and Theorem 4,

$$\begin{aligned}
\iota(\alpha f) &= \|\iota^*(\alpha f^*)\|_{\tilde{\mathcal{B}}^*} \cdot \text{PO}_{\tilde{\mathsf{B}}}(\iota^*(\alpha f^*)) \tag{87}\\
&= |\alpha| \, \|\iota^*(f^*)\|_{\tilde{\mathcal{B}}^*} \cdot \text{PO}_{\tilde{\mathsf{B}}}(\alpha \iota^*(f^*)). \tag{88}
\end{aligned}$$

By the symmetry of $\tilde{\mathsf{B}}$,

$$\begin{aligned}
\iota(\alpha f) &= |\alpha| \, \|\iota^*(f^*)\|_{\tilde{\mathcal{B}}^*} \cdot \text{sign}(\alpha) \, \text{PO}_{\tilde{\mathsf{B}}}(\iota^*(f^*)) \tag{89}\\
&= \alpha \, \iota(f). \tag{90}
\end{aligned}$$

Finally we show $\iota(f_1 + f_2) = \iota(f_1) + \iota(f_2)$ for all $f_1, f_2 \in \mathcal{B}$. Observe

$$\begin{aligned}
&\langle \iota(f_1) + \iota(f_2), \iota^*((f_1 + f_2)^*) \rangle \tag{91}\\
&= \langle \iota(f_1), \iota^*((f_1 + f_2)^*) \rangle + \langle \iota(f_2), \iota^*((f_1 + f_2)^*) \rangle \tag{92}\\
&= \langle f_1; (f_1 + f_2)^* \rangle + \langle f_2; (f_1 + f_2)^* \rangle \tag{93}\\
&= \langle f_1 + f_2; (f_1 + f_2)^* \rangle. \tag{94}
\end{aligned}$$

Therefore

$$\langle v, \iota^*((f_1 + f_2)^*) \rangle = \left\langle \frac{f_1 + f_2}{\|f_1 + f_2\|_{\mathcal{B}}}; (f_1 + f_2)^* \right\rangle, \tag{95}$$

$$\text{where} \quad v = \frac{\iota(f_1) + \iota(f_2)}{\|f_1 + f_2\|_{\mathcal{B}}}. \tag{96}$$

We now show $\|v\|_{\tilde{\mathcal{B}}} = 1$, which is equivalent to

$$\|\iota(f_1) + \iota(f_2)\|_{\tilde{\mathcal{B}}} = \|f_1 + f_2\|_{\mathcal{B}}. \tag{97}$$

Indeed, this can be easily seen from

$$\begin{aligned}
\text{LHS} &= \sup_{u \in \tilde{\mathsf{B}}^*} \langle \iota(f_1) + \iota(f_2), u \rangle \tag{98}\\
&= \sup_{g^* \in \mathsf{B}^*} \langle \iota(f_1) + \iota(f_2), \iota^*(g^*) \rangle \quad \text{(Lemma 4)} \tag{99}\\
&= \sup_{g^* \in \mathsf{B}^*} \langle f_1 + f_2; g^* \rangle \quad \text{(by Lemma 2)} \tag{100}\\
&= \text{RHS}. \tag{101}
\end{aligned}$$

By Lemma 5,

$$\max_{v \in \tilde{\mathsf{B}}} \langle v, \iota^*((f_1 + f_2)^*) \rangle = \max_{f \in \mathsf{B}} \langle f; (f_1 + f_2)^* \rangle. \tag{102}$$

Since the right-hand side is optimized at $f = (f_1 + f_2) / \|f_1 + f_2\|_{\mathcal{B}}$, we can see from (95) and $\|v\|_{\tilde{\mathcal{B}}} = 1$ that $v = \text{PO}_{\tilde{\mathsf{B}}}(\iota^*((f_1 + f_2)^*))$. Finally by definition (23), we conclude

$$\begin{aligned}
\iota(f_1 + f_2) &= \|\iota^*((f_1 + f_2)^*)\|_{\tilde{\mathcal{B}}^*} \cdot \text{PO}_{\tilde{\mathsf{B}}}(\iota^*((f_1 + f_2)^*)) \tag{103}\\
&= \|f_1 + f_2\|_{\mathcal{B}} \, v \quad \text{(by Lemma 1)} \tag{104}\\
&= \iota(f_1) + \iota(f_2).
\end{aligned}$$
□

Proof of Theorem 7. We assume that the kernel k is smooth and the function

$$z_{ij}(\lambda) = \tfrac{\partial}{\partial \lambda} k((\tilde{x}_\lambda, \tilde{y}_\lambda), (\cdot, \cdot)).$$

is in L_p so that R_{ij} is well-defined and finite-valued.

Clearly, using the representer theorem we can rewrite

$$R_{ij}(f) = \| \langle f, z_{ij}(\lambda) \rangle_{\mathcal{H}} \|_p. \tag{105}$$

Thus, R_{ij} is the composition of the linear map $f \mapsto g(\lambda; f) := \langle f, z_{ij}(\lambda) \rangle_{\mathcal{H}}$ and the L_p norm $g \mapsto \|g(\lambda)\|_p$. It follows from the chain rule that R_{ij} is convex, absolutely homogeneous, and Gâteaux differentiable (recall that the L_p norm is Gâteaux differentiable for $p \in (1, \infty)$). □

B. Analysis under Inexact Euclidean Embedding

We first rigorously quantify the inexactness in the Euclidean embedding T: $\mathcal{H} \to \mathbb{R}^d$, where $Tf = \tilde{f}$. To this end, let us consider a subspace based embedding, such as Nyström approximation. Here let T satisfy that there exists a countable set of orthonormal bases $\{e_i\}_{i=1}^{\infty}$ of $\mathcal{H}$, such that

1. $Te_k = 0$ for all $k > d$,

2. $\langle Tf, Tg \rangle = \langle f, g \rangle_{\mathcal{H}}, \;\; \forall f, g \in V := \text{span}\{e_1, \ldots, e_d\}$.

Clearly the Nyström approximation in (20) satisfies these conditions, where $d = n$, and $\{e_1, \ldots, e_d\}$ is any orthornormal basis of $\{k_{z_1}, \ldots, k_{z_d}\}$ (assuming d is no more than the dimensionality of $\mathcal{H}$).

As an immediate consequence, $\{Te_1, \ldots, Te_d\}$ forms an orthonormal basis of $\mathbb{R}^d$: $\langle Te_i, Te_j \rangle = \langle e_i, e_j \rangle_{\mathcal{H}} = \delta_{ij}$ for all $i, j \in [d]$. Besides, T is contractive because for all $f \in \mathcal{F}$,

$$\|Tf\|^2 = \left\| \sum_{i=1}^{d} \langle f, e_i \rangle_{\mathcal{H}} Te_i \right\|^2 \tag{106}$$

$$= \sum_{i=1}^{d} \langle f, e_i \rangle_{\mathcal{H}}^2 \le \|f\|_{\mathcal{H}}^2 . \tag{107}$$

By Definition 5, obviously k_{z_i} is 0-approximable under the Nyström approximation. If both f and g are ϵ-approximable, then $f + g$ must be (2ϵ)-approximable.

Lemma 6. *Let $f \in \mathcal{H}$ be ϵ-approximable by T, then for all $u \in \mathcal{H}$,*

$$|\langle u, f \rangle_{\mathcal{H}} - \langle Tu, Tf \rangle| \le \epsilon \|u\|_{\mathcal{H}} . \tag{108}$$

Proof. Let $f = \sum_{i=1}^{\infty} \alpha_i e_i$ and $u = \sum_{i=1}^{\infty} \beta_i e_i$. Then

$$|\langle u, f \rangle_{\mathcal{H}} - \langle Tu, Tf \rangle| \tag{109}$$

$$= \left| \sum_{i=1}^{\infty} \alpha_i \beta_i - \left\langle \sum_{i=1}^{d} \alpha_i Te_i, \sum_{j=1}^{d} \beta_j Te_j \right\rangle \right| \tag{110}$$

$$= \left| \sum_{i=d+1}^{\infty} \alpha_i \beta_i \right| \tag{111}$$

$$\le \left(\sum_{i=d+1}^{\infty} \alpha_i^2 \right)^{1/2} \left(\sum_{j=d+1}^{\infty} \beta_j^2 \right)^{1/2} \tag{112}$$

$$\le \epsilon \|u\|_{\mathcal{H}} .$$

□

Proof of Theorem 6. We first prove (30). Note for any $u \in \mathcal{F}$,

$$\langle u; g^* \rangle = [u, g] \tag{113}$$

$$= \lim_{t \to 0} \frac{1}{2} \left[\|tu + g\|_{\mathcal{B}}^2 - \|g\|_{\mathcal{B}}^2 \right] \tag{114}$$

$$= \left\langle u, g + \nabla R^2(g) \right\rangle_{\mathcal{H}} . \tag{115}$$

The differentiability of R^2 is guaranteed by the Gâteaux differentiability. Letting $g^* = \sum_i \alpha_i G_{v_i}^*$, it follows that

$$\langle u; g^* \rangle = \sum_i \alpha_i u(v_i) = \left\langle u, \sum_i \alpha_i k_{v_i} \right\rangle_{\mathcal{H}} . \tag{116}$$

So $\sum_i \alpha_i k_{v_i} = g + \nabla R^2(g)$, and by the definition of ι^*

$$\iota^*(g^*) = \sum_i \alpha_i T k_{v_i} = T a_g \tag{117}$$

$$\text{where} \quad a_g := \sum_i \alpha_i k_{v_i} = g + \nabla R^2(g). \tag{118}$$

Similarly,

$$\iota^*(f^*) = T a_f, \quad \text{where} \quad a_f := f + \nabla R^2(f). \tag{119}$$

By assumption $\arg\max_{h \in S} \langle h, g \rangle_{\mathcal{H}}$ is ϵ-approximable, and hence a_g is $O(\epsilon)$-approximable. Similarly, a_f is also $O(\epsilon)$-approximable.

Now let us consider

$$v^\circ := \arg \max_{v \in \mathbb{R}^d : \|v\|^2 + \sup_{h \in S} \langle v, Th \rangle^2 \le 1} \langle v, T a_f \rangle \tag{120}$$

$$u^\circ := \arg \max_{u \in \mathcal{F} : \|u\|_{\mathcal{H}}^2 + \sup_{h \in S} \langle u, h \rangle_{\mathcal{H}}^2 \le 1} \langle u, a_f \rangle_{\mathcal{H}} . \tag{121}$$

By definition, $\iota(f) = v^\circ$. Also note that $u^\circ = f$ because $\langle u, a_f \rangle_{\mathcal{H}} = \langle u; f^* \rangle$ for all $u \in \mathcal{F}$. We will then show that

$$\|\iota(f) - Tf\| = \|v^\circ - Tu^\circ\| = O(\sqrt{\epsilon}), \tag{122}$$

which allows us to derive that

$$\langle f; g^* \rangle = \langle f, a_g \rangle_{\mathcal{H}} \tag{123}$$

$$= \langle Tf, Ta_g \rangle + O(\epsilon) \quad \text{(by Lemma 6)} \tag{124}$$

$$= \langle Tu^\circ, Ta_g \rangle + O(\epsilon) \tag{125}$$

$$= \langle v^\circ, Ta_g \rangle + O(\sqrt{\epsilon}) \quad \text{(by (122))} \tag{126}$$

$$= \langle \iota(f), \iota^*(g^*) \rangle + O(\sqrt{\epsilon}). \quad \text{(by (117))} \tag{127}$$

Finally, we prove (122). Denote

$$w^\circ := \arg \max_{w \in \mathcal{F}: \|w\|^2_{\mathcal{H}} + \sup_{h \in S} \langle Tw, Th \rangle^2 \leq 1} \langle w, a_f \rangle_{\mathcal{H}} . \tag{128}$$

We will prove that $\|v^\circ - Tw^\circ\| = O(\epsilon^2)$ and $\|u^\circ - w^\circ\|_{\mathcal{H}} = O(\sqrt{\epsilon})$. They will imply (122) because by the contractivity of T, $\|T(u^\circ - w^\circ)\| \leq \|u^\circ - w^\circ\|_{\mathcal{H}}$.

Step 1: $\|v^\circ - Tw^\circ\| = O(\epsilon^2)$. Let $w = w_1 + w_2$ where $w_1 \in V$ and $w_2 \in V^\perp$. So $Tw = Tw_1$ and $\|Tw\| = \|w_1\|_{\mathcal{H}}$. Similarly decompose a_f as $a_1 + a_2$, where $a_1 = Ta_f \in V$ and $a_2 \in V^\perp$. Now the optimization over w becomes

$$\max_{w_1 \in V, w_2 \in V^\perp} \quad \langle w_1, a_1 \rangle_{\mathcal{H}} + \langle w_2, a_2 \rangle_{\mathcal{H}} \tag{129}$$

$$s.t. \quad \|w_1\|^2_{\mathcal{H}} + \|w_2\|^2_{\mathcal{H}} + \sup_{h \in S} \langle Tw_1, Th \rangle^2 \leq 1. \tag{130}$$

Let $\|w_2\|^2 = 1 - \alpha$ where $\alpha \in [0, 1]$. Then the optimal value of $\langle w_2, a_2 \rangle_{\mathcal{H}}$ is $\sqrt{1-\alpha}\, \|a_2\|_{\mathcal{H}}$. Since $\langle w_1, a_1 \rangle_{\mathcal{H}} = \langle Tw_1, Ta_1 \rangle$, the optimization over w_1 can be written as

$$\min_{w_1 \in V} \langle Tw_1, Ta_1 \rangle \tag{131}$$

$$s.t. \ \|Tw_1\|^2 + \sup_{h \in S} \langle Tw_1, Th \rangle^2 \leq \alpha. \tag{132}$$

Change variable by $v = Tw_1$. Then compare with the optimization of v in (120), and we can see that $v^\circ = Tw_1^\circ / \sqrt{\alpha}$. Overall the optimal objective value of (129) under $\|w_2\|^2 = 1 - \alpha$ is $\sqrt{1-\alpha}\, \|a_2\|_{\mathcal{H}} + \sqrt{\alpha} p$ where p is the optimal objective value of (120). So the optimal α is $\frac{p^2}{p^2 + \|a_2\|^2_{\mathcal{H}}}$, and hence

$$\|v^\circ - Tw^\circ\| = \|v^\circ - Tw_1^\circ\| = \left\|v^\circ - \sqrt{\alpha} v^\circ\right\| \tag{133}$$

$$= (1 - \sqrt{\alpha})\, \|v^\circ\| \leq 1 - \sqrt{\alpha}. \tag{134}$$

Since a_f is $O(\epsilon)$-approximable, so $\|a_2\|_{\mathcal{H}} = O(\epsilon)$ and

$$1 - \sqrt{\alpha} = \frac{1 - \alpha}{1 + \sqrt{\alpha}} = O(\|a_2\|^2_{\mathcal{H}}) = O(\epsilon^2). \tag{135}$$

Step 2: $\|u^\circ - w^\circ\|_{\mathcal{H}} = O(\sqrt{\epsilon})$. Motivated by Theorem 8, we consider two equivalent problems:

$$\hat{u}^\circ = \arg \max_{u \in \mathcal{F}: \langle u, a_f \rangle_{\mathcal{H}} = 1} \left\{ \|u\|^2_{\mathcal{H}} + \sup_{h \in S} \langle u, h \rangle^2_{\mathcal{H}} \right\} \tag{136}$$

$$\hat{w}^\circ = \arg \max_{w \in \mathcal{F}: \langle w, a_f \rangle_{\mathcal{H}} = 1} \left\{ \|w\|^2_{\mathcal{H}} + \sup_{h \in S} \langle Tw, Th \rangle^2 \right\} . \tag{137}$$

Again we can decompose u into $U := \text{span}\{a_f\}$ and its orthogonal space $U^\perp$. Since $\langle u, a_f \rangle_{\mathcal{H}} = 1$, the component of u in U must be $\bar{a}_f := a_f / \|a_f\|^2_{\mathcal{H}}$. So

$$\hat{u}^\circ = \bar{a}_f + \arg \max_{u^\perp \in U^\perp} \left\{ \left\|u^\perp\right\|^2_{\mathcal{H}} + \sup_{h \in S} \left\langle u^\perp + \bar{a}_f, h \right\rangle^2_{\mathcal{H}} \right\} . \tag{138}$$

Similarly,

$$w^\circ = \bar{a}_f + \arg \max_{w^\perp \in U^\perp} \Bigg\{ \left\|w^\perp\right\|^2_{\mathcal{H}} \tag{139}$$

$$+ \sup_{h \in S} \left\langle T(w^\perp + \bar{a}_f), Th \right\rangle^2_{\mathcal{H}} \Bigg\}. \tag{140}$$

We now compare the objective in the above two argmax forms. Since any $h \in S$ is ϵ-approximable, so for any $x \in \mathcal{F}$:

$$|\langle x, h \rangle_{\mathcal{H}} - \langle Tx, Th \rangle_{\mathcal{H}}| = O(\epsilon). \tag{141}$$

Therefore tying $u^\perp = w^\perp = x$, the objectives in the argmax of (138) and (139) differ by at most $O(\epsilon)$. Therefore their optimal objective values are different by at most $O(\epsilon)$. Since both objectives are (locally) strongly convex in $U^\perp$, the RKHS distance between the optimal $u^\perp$ and the optimal $w^\perp$ must be $O(\sqrt{\epsilon})$. As a result $\|\hat{u}^\circ - \hat{w}^\circ\|_{\mathcal{H}} = O(\sqrt{\epsilon})$.

Finally to see $\|u^\circ - w^\circ\|_{\mathcal{H}} = O(\epsilon)$, just note that by Theorem 8, u° and w° simply renormalize $\hat{u}^\circ$ and $\hat{w}^\circ$ to the unit sphere of $\|\cdot\|_{\mathcal{B}}$, respectively. So again $\|u^\circ - w^\circ\|_{\mathcal{H}} = O(\sqrt{\epsilon})$.

In the end, we prove (31). The proof of $\iota(\alpha f) = \alpha \iota(f)$ is exactly the same as that for Theorem 4. To prove (31), note that $f + g$ is (2ϵ)-approximable. Therefore applying (122) on f, g, $f + g$, we get

$$\|\iota(f) - Tf\| = O(\sqrt{\epsilon}), \tag{142}$$

$$\|\iota(fg) - Tg\| = O(\sqrt{\epsilon}), \tag{143}$$

$$\|\iota(f + g) - T(f + g)\| = O(\sqrt{\epsilon}). \tag{144}$$

Combining these three relations, we conclude (31). □

C. Solving the Polar Operator

Theorem 8. *Suppose J is continuous and $J(\alpha x) = \alpha^2 J(x) \geq 0$ for all x and $\alpha \geq 0$. Then x is an optimal solution to*

$$P: \quad \max_x a^\top x, \quad s.t. \quad J(x) \leq 1, \tag{145}$$

if, and only if, $J(x) = 1$, $c := a^\top x > 0$, and $\hat{x} := x/c$ is an optimal solution to

$$Q: \quad \min_x J(x), \quad s.t. \quad a^\top x = 1. \tag{146}$$

Proof. We first show the "only if" part. Since $J(0) = 0$ and J is continuous, the optimal objective value of P must be positive. Therefore $c > 0$. Also note the optimal x for P must satisfy $J(x) = 1$ because otherwise one can scale up x to increase the objective value of P. To show $\hat{x}$ optimizes Q, suppose otherwise there exists y such that

$$a^\top y = 1, \quad J(y) < J(\hat{x}). \tag{147}$$

Then letting

$$z = J(y)^{-1/2} y, \tag{148}$$

we can verify that

$$J(z) = 1, \tag{149}$$

$$a^\top z = J(y)^{-1/2} a^\top y = J(y)^{-1/2} \tag{150}$$

$$> J(\hat{x})^{-1/2} = c J(x)^{-1/2} = c = a^\top x. \tag{151}$$

So z is a feasible solution for P, and is strictly better than x. Contradiction.

We next show the "if" part: for any x, if $J(x) = 1$, $c := a^\top x > 0$, and $\hat{x} := x/c$ is an optimal solution to Q, then x must optimize P. Suppose otherwise there exists y, such that $J(y) \leq 1$ and $a^\top y > a^\top x > 0$. Then consider $z := y/a^\top y$. It is obviously feasible for Q, and

$$J(z) = (a^\top y)^{-2} J(y) < (a^\top x)^{-2} J(y) \tag{152}$$

$$\leq (a^\top x)^{-2} J(x) = J(\hat{x}). \tag{153}$$

This contradicts with the optimality of $\hat{x}$ for Q. □

Projection to hyperplane To solve problem (28), we use LBFGS with each step projected to the feasible domain, a hyperplane. This requires solving, for given c and a,

$$\min_x \frac{1}{2} \|x - c\|^2, \quad s.t. \quad a^\top x = 1. \tag{154}$$

Write out its Lagrangian and apply strong duality thanks to convexity:

$$\min_x \max_\lambda \frac{1}{2} \|x - c\|^2 - \lambda(a^\top x - 1) \tag{155}$$

$$= \max_\lambda \min_x \frac{1}{2} \|x - c\|^2 - \lambda(a^\top x - 1) \tag{156}$$

$$= \max_\lambda \frac{1}{2} \lambda^2 \|a\|^2 - \lambda^2 \|a\|^2 - \lambda a^\top c + \lambda, \tag{157}$$

where $x = c + \lambda a$. The last step has optimal

$$\lambda = (1 - a^\top c) / \|a\|^2 . \tag{158}$$

D. Gradient in Dual Coefficients

We first consider the case where S is a finite set, and denote as z_i the RKHS Nyström approximation of its i-th element. When f^* has the form of (12), we can compute $\iota(f)$ by using the Euclidean counterpart of Theorem 3 as follows:

$$\arg\max_u \; u^\top \sum\nolimits_j c_j k_j \tag{159}$$

$$s.t. \; \|u\|^2 + (z_i^\top u)^2 \leq 1, \quad \forall\, i, \tag{160}$$

where k_j the the Nyström approximation of $k(x_j, \cdot)$.

Writing out the Lagrangian with dual variables λ_i:

$$u^\top \sum_j c_j k_j + \sum_i \lambda_i \left(\|u\|^2 + (z_i^\top u)^2 - 1 \right), \tag{161}$$

we take derivative with respect to u:

$$X^\top c + 2\mathbf{1}^\top \lambda u + 2 Z \Lambda Z^\top u = 0. \tag{162}$$

where $X = (k_1, k_2, \ldots)$, $Z = (z_1, z_2, \ldots)$, $\lambda = (\lambda_1, \lambda_2, \ldots)$, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots)$ (diagonal matrix), and $\mathbf{1}$ is a vector of all ones. This will hold for $c + \Delta_c$, $\lambda + \Delta_\lambda$ and $u + \Delta_u$:

$$X^\top (c + \Delta_c) + 2\mathbf{1}^\top (\lambda + \Delta_\lambda)(u + \Delta u) \tag{163}$$

$$+ 2Z(\Lambda + \Delta_\Lambda) Z^\top (u + \Delta_u) = 0. \tag{164}$$

Subtract it by (162), we obtain

$$X^\top \Delta_c + 2(\mathbf{1}^\top \Delta_\lambda) u + 2(\mathbf{1}^\top \lambda) \Delta_u \tag{165}$$

$$+ 2 Z \Delta_\Lambda Z^\top u + 2 Z \Lambda Z^\top \Delta_u = 0. \tag{166}$$

The complementary slackness writes

$$\lambda_i (\|u\|^2 + (z_i^\top u)^2 - 1) = 0. \tag{167}$$

This holds for $\lambda + \Delta_\lambda$ and $u + \Delta_u$:

$$(\lambda_i + \Delta_{\lambda_i})(\|u + \Delta_u\|^2 + (z_i^\top u + z_i^\top \Delta_u)^2 - 1) = 0. \tag{168}$$

Subtract it by (167), we obtain

$$\Delta_{\lambda_i}(\|u\|^2 + (z_i^\top u)^2 - 1) + 2\lambda_i(u + (z_i^\top u)z_i)^\top \Delta_u = 0. \tag{169}$$

Putting together (165) and (169), we obtain

$$S\begin{pmatrix}\Delta_u \\ \Delta_\lambda\end{pmatrix} = \begin{pmatrix}-X^\top \Delta_c \\ 0\end{pmatrix}, \tag{170}$$

where S is

$$\begin{pmatrix} 2(\mathbf{1}^\top\lambda)I + 2Z\Lambda Z^\top & 2u\mathbf{1}^\top + 2Z\,\mathrm{diag}(Z^\top u) \\ 2\Lambda(\mathbf{1}u^\top + \mathrm{diag}(Z^\top u)Z^\top) & \mathrm{diag}(\|u\|^2 + (z_i^\top u)^2 - 1) \end{pmatrix}. \tag{171}$$

Therefore

$$\frac{\mathrm{d}u}{\mathrm{d}c} = \begin{pmatrix} I & 0 \end{pmatrix} S^{-1} \begin{pmatrix}-X^\top \\ 0\end{pmatrix}. \tag{172}$$

Finally we investigate the case when S is not finite. In such a case, the elements z in S that attain $\|u\|^2 + (z^\top u)^2 = 1$ for the optimal u are still finite in general. For all other z, the complementary slackness implies the corresponding λ element is 0. As a result, the corresponding diagonal entry in the bottom-right block of S is nozero, while the corresponding row in the bottom-left block of S is straight 0. So the corresponding entry in Δ_λ in (170) plays no role, and can be pruned. In other words, all $z \in S$ such that $\|u\|^2 + (z^\top u)^2 < 1$ can be treated as nonexistent.

The emprirical loss depends on $f(x_j)$, which can be computed by $\iota(f)^\top k_j$. Since $\iota(f) = (u^\top \sum_j c_j k_j)u$, (172) allows us to backpropagate the gradient in $\iota(f)$ into the grdient in $\{c_j\}$.

E. Experiments

E.1. Additional experimental results on mixup

Results. We first present more detailed experimental results for the mixup learning. Following the algorithms described in Section 7.2, each setting was evaluated 10 times with randomly sampled training and test data. The mean and standard deviation are reported in Table 2. Since the results of Embed and Vanilla have the smallest difference under $n = 1000, p = 4n$, for each dataset, we show scatter plots of test accuracy under 10 runs for this setting. In Figure 2, the x-axis represents accuracy of Embed method, and the y-axis represents the accuracy of Vanilla. Obviously, most points fall above the diagonal, meaning Embed method outperforms Vanilla most of the time.

Visualization. To show that Embed learned better representations in mixup, we next visualized the impact of the two different methods. Figure 3 plots how the loss value of three randomly sampled pairs of test examples changes as a function of λ in (32). Each subplot here corresponds to a randomly chosen pair. By increasing λ from 0 to 1 with a step size 0.1, we obtained different mixup representations. We then applied the trained classifiers on these representations to compute the loss value. As shown in Figure 3, Embed always has a lower loss, especially when Vanilla is at its peak loss value. Recall in (33), Embed learns representations by considering the λ that maximizes the change; this figure exactly verified this behavior and Embed learns better representation.

E.2. Additional experiments for structured multilabel prediction

Here, we provide more detailed results for our method applied to structured multilabel prediction, as described in Section 6.

Accuracy on multiple runs. We repeated the experiment, detailed in Section 7.3 and tabulated in Table 3 ten times for all the three algorithms. Figures 4,5,6 show the accuracy plot of our method (Embed) compared with baselines (ML-SVM and HR-SVM) on Enron (Klimt and Yang, 2004), WIPO (Rousu et al., 2006), Reuters (Lewis et al., 2004) datasets with $100/100, 200/200, 500/500$ randomly drawn train/test examples over 10 runs.

Comparing constraint violations. In this experiment, we demonstrate the effectiveness of the model's ability to embed structures explicitly. Recall that for the structured multilabel prediction task, we wanted to incorporate two types of constraints (i) *implication*, (ii) *exclusion*. To test if our model (Embed) indeed learns representations that respect these constraints, we counted the number of test examples that violated the implication and exclusion constraints from the predictions. We repeated the test for ML-SVM and HR-SVM.

We observed that HR-SVM and Embed successfully modeled implications on all the datasets. This is not surprising as HR-SVM takes the class hierarchy into account. The exclusion constraint, on the other hand, is a "derived" constraint and is not directly modeled by HR-SVM. Therefore, on datasets where Embed performed significantly better than HR-SVM, we might expect fewer exclusion violations by Embed compared to HR-SVM. To verify this intuition, we considered the Enron dataset with $200/200$ train/test split where Embed performed better than HR-SVM. The constraint violations are shown as a line plot in Figure 7, with the constraint index on the x-axis and number of examples violating the constraint on the y-axis.

Recall again that predictions in Embed for multilabel prediction are made using a linear classifier. Therefore the superior performance of Embed in this case, can be attributed to accurate representations learned by the model.

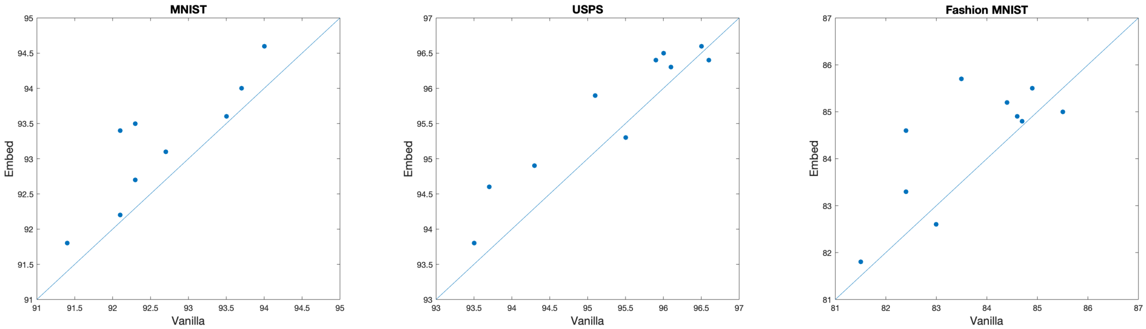


Figure 2: Scatter plot of test accuracy for mixup: $n = 1000$, $p = 4n$

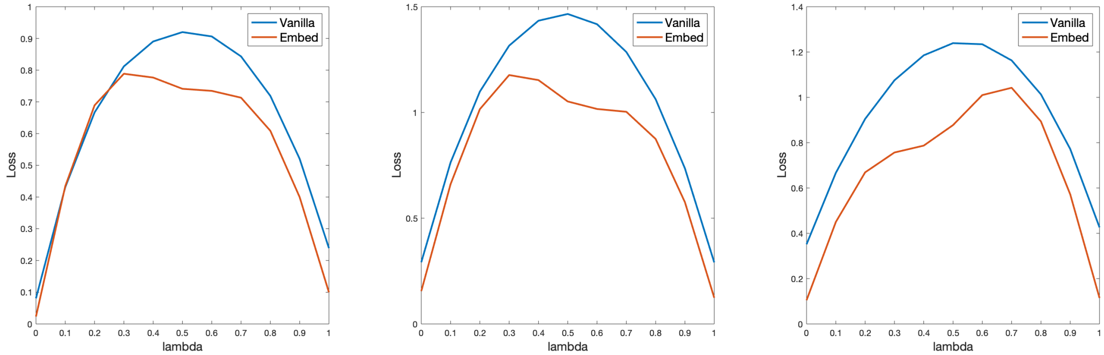
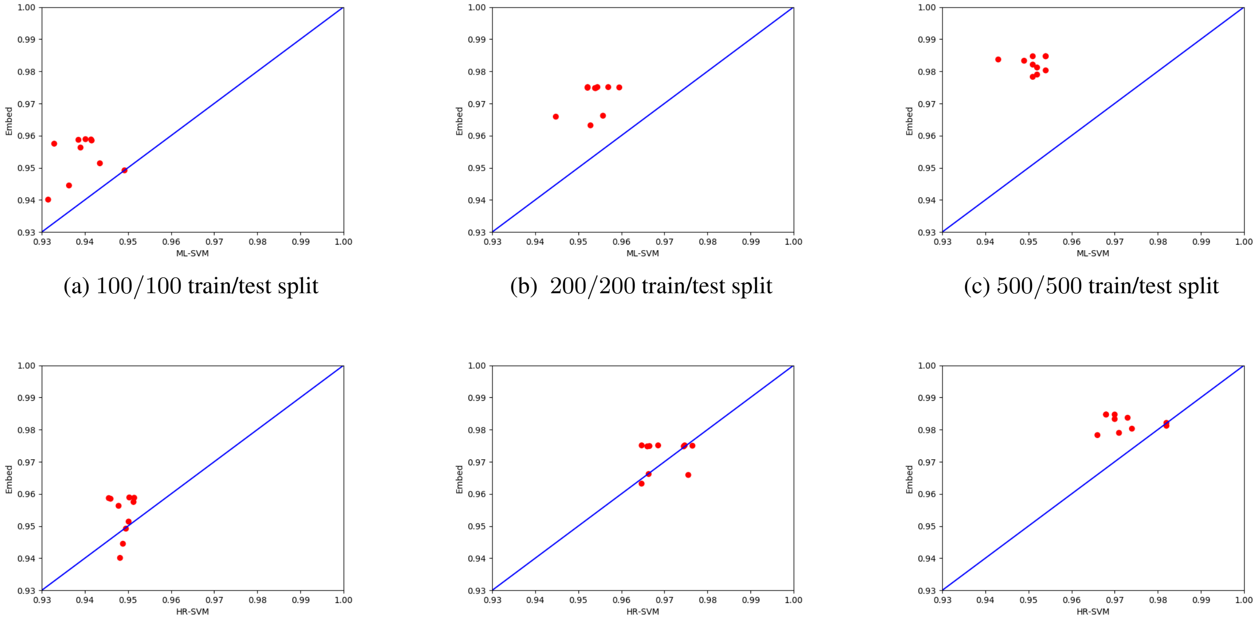


Figure 3: Plots of three different pairs of test examples, showing how loss values change as a function of λ



(a) 100/100 train/test split

(b) 200/200 train/test split

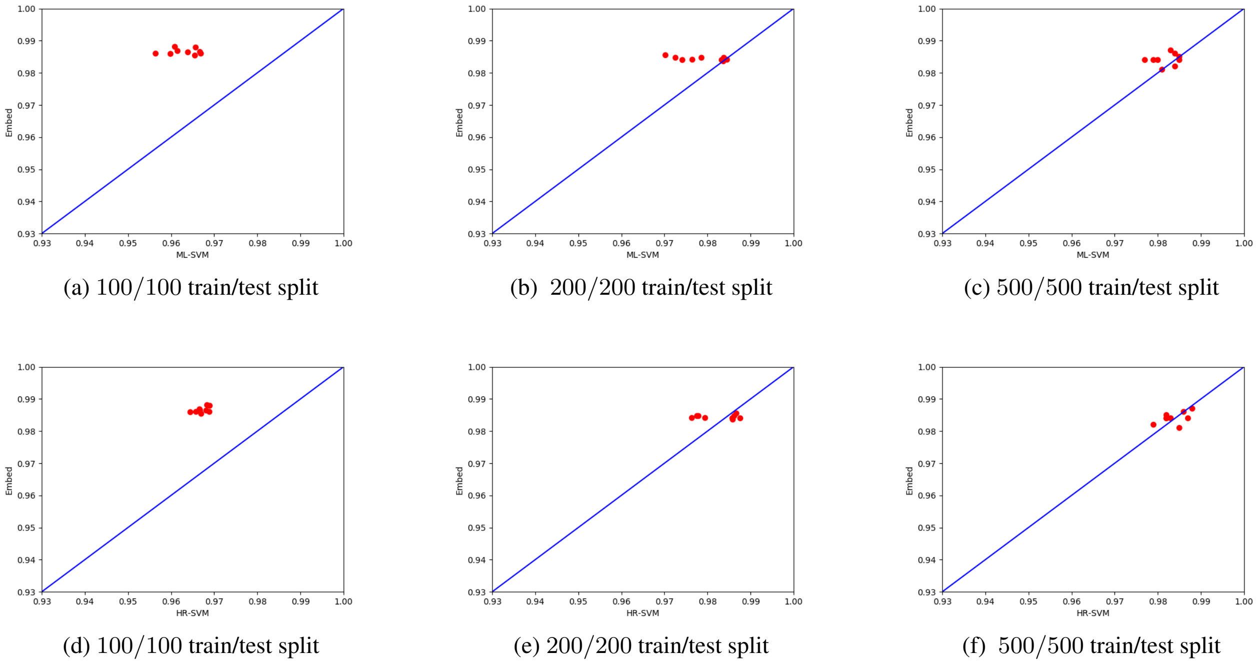
(c) 500/500 train/test split

(d) 100/100 train/test split

(e) 200/200 train/test split

(f) 500/500 train/test split

Figure 4: Test accuracy of ML-SVM vs Embed (top row) and HR-SVM vs Embed (bottom row) 10 runs on the Reuters dataset



(a) 100/100 train/test split

(b) 200/200 train/test split

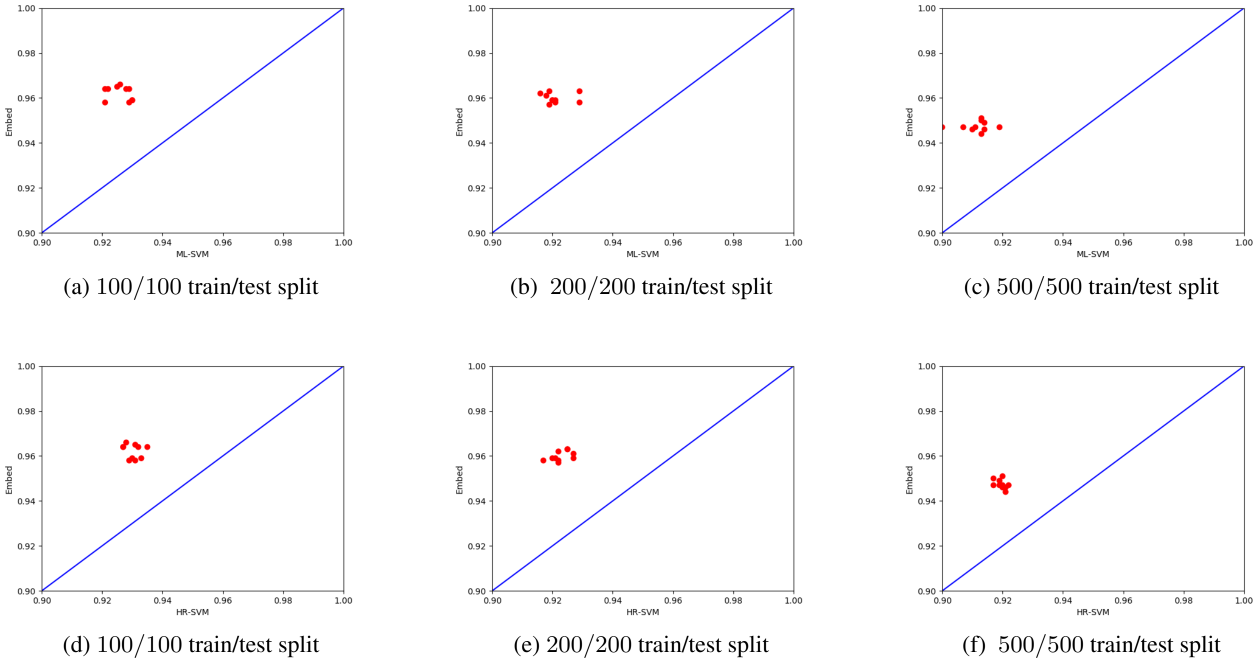
(c) 500/500 train/test split

(d) 100/100 train/test split

(e) 200/200 train/test split

(f) 500/500 train/test split

Figure 5: Test accuracy of ML-SVM vs Embed (top row) and HR-SVM vs Embed (bottom row) 10 runs on the WIPO dataset



(a) 100/100 train/test split

(b) 200/200 train/test split

(c) 500/500 train/test split

(d) 100/100 train/test split

(e) 200/200 train/test split

(f) 500/500 train/test split

Figure 6: Test accuracy of ML-SVM vs Embed (top row) and HR-SVM vs Embed (bottom row) 10 runs on the ENRON dataset

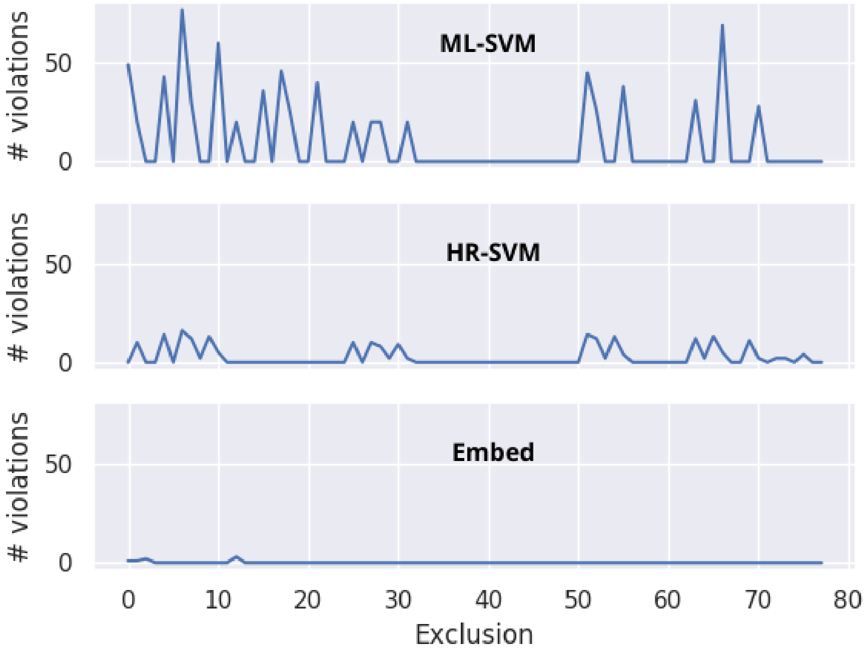


Figure 7: The number of violations for each exclusion constraint on the test set by (from top) ML-SVM, HR-SVM, and Embed on the Enron dataset with $200/200$ train/test examples.