A Supplementary Material

A.1 Mathematical Results

Corollary 2. *For $t > 0$ and a prior distribution p over $\mathbb{R}^d$, with probability $1-\varepsilon$ over the choice of $S \sim \mathcal{D}^n$, we have for all q on $\mathbb{R}^d$:*

$$\mathcal{L}_{\mathcal{D}}(k_q) \leq \widehat{\mathcal{L}}_S(k_q) + \frac{2}{t}\left(\mathrm{KL}(q\|p) + \frac{t^2}{2(n-1)} + \ln\frac{n+1}{\varepsilon}\right).$$

Proof. We want to bound

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(k_q) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\ \mathop{\mathbf{E}}_{(\mathbf{x}',y')\sim\mathcal{D}}\ \mathop{\mathbf{E}}_{\boldsymbol{\omega}\sim q} \ell\Big(h_{\boldsymbol{\omega}}(\mathbf{x}-\mathbf{x}'), \lambda(y,y')\Big)\\ &= \mathop{\mathbf{E}}_{(\mathbf{x}',y')\sim\mathcal{D}} \mathcal{L}'_{\mathcal{D}}(k_q)\,,\end{aligned}$$

where $\mathcal{L}'_{\mathcal{D}}(k_q)$ is the alignment loss of the kernel k_q centered on $(\mathbf{x}', y') \sim \mathcal{D}$ (see Equation (10)).

Let $t > 0$ and p a distribution on $\mathbb{R}^d$. By applying the PAC-Bayesian theorem, with $\varepsilon_0 \in (0,1)$, we have

$$\Pr_{S\sim\mathcal{D}^n}\left(\forall q \text{ on } \mathbb{R}^d : \mathcal{L}_{\mathcal{D}}(k_q) \leq \frac{1}{n}\sum_{i=1}^n \mathcal{L}^i_{\mathcal{D}}(k_q) + \frac{1}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2n} + \ln\frac{1}{\varepsilon_0}\right]\right) \geq 1-\epsilon_0\,.$$

Moreover, we have that for each $i \in \{1,\ldots,n\}$, with a $\varepsilon_i \in (0,1)$, we have

$$\Pr_{S\sim\mathcal{D}^n}\left(\forall q \text{ on } \mathbb{R}^d : \mathcal{L}^i_{\mathcal{D}}(k_q) \leq \widehat{\mathcal{L}}^i_S(k_q) + \frac{1}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2(n-1)} + \ln\frac{1}{\varepsilon_i}\right]\right) \geq 1-\epsilon_i\,.$$

By combining above probabilistic results with $\varepsilon_0 = \varepsilon_1 = \cdots = \varepsilon_n = \frac{\varepsilon}{n+1}$, we obtain that, with probability at least $1-\varepsilon$,

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(k_q) &= \mathop{\mathbf{E}}_{(\mathbf{x}',y')\sim\mathcal{D}} \mathcal{L}'_{\mathcal{D}}(k_q)\\ &\leq \frac{1}{n}\sum_{i=1}^n\left[\widehat{\mathcal{L}}^i_S(k_q) + \frac{1}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2(n-1)} + \ln\frac{n+1}{\varepsilon}\right]\right] + \frac{1}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2n} + \ln\frac{n+1}{\varepsilon}\right]\\ &= \widehat{\mathcal{L}}_S(k_q) + \frac{1}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2(n-1)} + \ln\frac{n+1}{\varepsilon}\right] + \frac{1}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2n} + \ln\frac{n+1}{\varepsilon}\right]\\ &\leq \widehat{\mathcal{L}}_S(k_q) + \frac{2}{t}\left[\mathrm{KL}(q\|p) + \frac{t^2}{2(n-1)} + \ln\frac{n+1}{\varepsilon}\right].\end{aligned}$$

□

Lemma 6. *For any data-generating distribution $\mathcal{D}$:*

$$\mathop{\mathbf{Var}}_{S'\sim\mathcal{D}^n}\left(\mathcal{L}_{S'}(h_{\boldsymbol{\omega}})\right) \leq \frac{1}{4n}\,.$$

Proof. Given $S' = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$, we denote

$$\mathcal{F}_{\boldsymbol{\omega}}(S') \coloneqq \mathcal{F}_{\boldsymbol{\omega}}\big((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\big) \coloneqq \mathcal{L}_{S'}(h_{\boldsymbol{\omega}}) = \frac{1}{n(n-1)}\sum_{i\neq j}^n \ell\Big(h_{\boldsymbol{\omega}}(\mathbf{x}_i-\mathbf{x}_j), \lambda(y_i,y_j)\Big)\,.$$

The function $\mathcal{F}_{\boldsymbol{\omega}}$ above has the *bounded differences property*. That is, for each $i \in \{1,\ldots,n\}$:

$$\sup_{S',\mathbf{x}^*\in\mathbb{R}^d,y^*\in Y}\left|\mathcal{F}_{\boldsymbol{\omega}}\big((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\big) - \mathcal{F}_{\boldsymbol{\omega}}\big((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_{i-1},y_{i-1}),(\mathbf{x}^*,y^*),(\mathbf{x}_{i+1},y_{i+1}),\ldots,(\mathbf{x}_n,y_n)\big)\right| \leq \frac{1}{n},$$

Thus, we apply the Efron-Stein inequality (following Boucheron et al., 2013, Corollary 3.2) to obtain

$$\underset{S'\sim\mathcal{D}^n}{\mathbf{Var}}\left(\mathcal{F}_{\boldsymbol{\omega}}(S')\right) \leq \frac{1}{4}\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 = \frac{1}{4n}\,.$$

□

A.2 Kernel Alignment Loss Computation

The kernel learning algorithms presented in Section 5 require to compute the empirical kernel alignment loss for each hypothesis $h_{\boldsymbol{\omega}}$, given by

$$\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}}) = \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\ell(h_{\boldsymbol{\omega}}(\boldsymbol{\delta}_{ij}),\lambda_{ij})\,. \tag{21}$$

A naive implementation of Equation (21) would need $O(n^2)$ steps. Propositions 7 and 8 below show how to rewrite Equation (21) in a form that needs $O(n)$ steps. Proposition 7 is dedicated to the binary classification, and is equivalent to the computation method proposed by Sinha and Duchi (2016). By Proposition 8, we extend the result to the multi-classification case.

Proposition 7 (Binary classification). *When $S = (\mathbf{x}_i, y_i)_{i=1}^n \in (\mathbb{R}^d \times \{-1,1\})^n$, we have*

$$\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}}) = \frac{n}{2(n-1)} - \frac{1}{2n(n-1)}\left[\left(\sum_{i=1}^{n} y_i\cos(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2 + \left(\sum_{i=1}^{n} y_i\sin(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2\right].$$

That is, in the binary classification case ($y\in\{-1,1\}$), one can compute the empirical alignment loss $\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}})$ in $O(n)$ steps.

Proof. Using the cosine trigonometric identity

$$\begin{aligned}
\sum_{i\neq j}^{n}\lambda_{ij}h_{\boldsymbol{\omega}}(\mathbf{x}_i-\mathbf{x}_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_iy_j\cos(\boldsymbol{\omega}\cdot(\mathbf{x}_i-\mathbf{x}_j)) - n\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} y_iy_j\left(\cos(\boldsymbol{\omega}\cdot\mathbf{x}_i)\cos(\boldsymbol{\omega}\cdot\mathbf{x}_j) + \sin(\boldsymbol{\omega}\cdot\mathbf{x}_i)\sin(\boldsymbol{\omega}\cdot\mathbf{x}_j)\right) - n\\
&= \left(\sum_{i=1}^{n} y_i\cos(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2 + \left(\sum_{i=1}^{n} y_i\sin(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2 - n
\end{aligned}$$

Thus,

$$\begin{aligned}
\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}}) &= \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\ell(h_{\boldsymbol{\omega}}(\boldsymbol{\delta}_{ij}),\lambda_{ij})\\
&= \frac{1}{n(n-1)}\sum_{i\neq j}^{n}\frac{1-\lambda_{ij}h_{\boldsymbol{\omega}}(\boldsymbol{\delta}_{ij})}{2}\\
&= \frac{1}{2} - \frac{1}{2n(n-1)}\sum_{i\neq j}^{n}\lambda_{ij}h_{\boldsymbol{\omega}}(\mathbf{x}_i-\mathbf{x}_j)\\
&= \frac{1}{2} - \frac{1}{2n(n-1)}\left[\left(\sum_{i=1}^{n} y_i\cos(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2 + \left(\sum_{i=1}^{n} y_i\sin(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2 - n\right]\\
&= \frac{n}{2(n-1)} - \frac{1}{2n(n-1)}\left[\left(\sum_{i=1}^{n} y_i\cos(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2 + \left(\sum_{i=1}^{n} y_i\sin(\boldsymbol{\omega}\cdot\mathbf{x}_i)\right)^2\right].
\end{aligned}$$

□

Proposition 8 (Multi-class classification)**.** *When* $S = (\mathbf{x}_i, y_i)_{i=1}^n \in (\mathbb{R}^d \times \{1, \dots, L\})^n$*, we have*

$$\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}}) = \frac{n}{2(n-1)} - \frac{1}{2n(n-1)} \left[2 \sum_{y=1}^{L} (c_y^2 + s_y^2) - \left(\sum_{y=1}^{L} c_y \right)^2 - \left(\sum_{y=1}^{L} s_y \right)^2 \right],$$

with

$$c_y \coloneqq \sum_{\mathbf{x} \in S_y} \cos(\boldsymbol{\omega} \cdot \mathbf{x}) \qquad \textit{and} \qquad s_y \coloneqq \sum_{\mathbf{x} \in S_y} \sin(\boldsymbol{\omega} \cdot \mathbf{x}) \,.$$

That is, in the multi-class classification case with L classes ($y \in \{1, \dots, L\})^n$), one can compute the empirical alignment loss $\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}})$ in $O(n)$ steps.

Proof.

$$\begin{aligned}
\sum_{i \neq j}^{n} \lambda_{ij} h_{\boldsymbol{\omega}}(\mathbf{x}_i - \mathbf{x}_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij} \cos(\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j)) - n \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (2I[y_i = y_j] - 1) \cos(\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j)) - n \\
&= 2 \sum_{i=1}^{n} \sum_{j=1}^{n} I[y_i = y_j] \cos(\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j)) - \sum_{i=1}^{n} \sum_{j=1}^{n} \cos(\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j)) - n
\end{aligned}$$

Let's denote $S_y \coloneqq \{\mathbf{x}_i | (\mathbf{x}_i, y) \in S\}$. We have

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} I[y_i = y_j] \cos(\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j)) &= \sum_{y=1}^{L} \sum_{\mathbf{x} \in S_y} \sum_{\mathbf{x}' \in S_y} \cos(\boldsymbol{\omega} \cdot (\mathbf{x} - \mathbf{x}')) \\
&= \sum_{y=1}^{L} \left[\left(\sum_{\mathbf{x} \in S_y} \cos(\boldsymbol{\omega} \cdot \mathbf{x}) \right)^2 + \left(\sum_{\mathbf{x} \in S_y} \sin(\boldsymbol{\omega} \cdot \mathbf{x}) \right)^2 \right],
\end{aligned}$$

and

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \cos(\boldsymbol{\omega} \cdot (\mathbf{x}_i - \mathbf{x}_j)) = \left(\sum_{y=1}^{L} \sum_{\mathbf{x} \in S_y} \cos(\boldsymbol{\omega} \cdot \mathbf{x}) \right)^2 + \left(\sum_{y=1}^{L} \sum_{\mathbf{x} \in S_y} \sin(\boldsymbol{\omega} \cdot \mathbf{x}) \right)^2 .$$

Thus, we can rewrite

$$\sum_{i \neq j}^{n} \lambda_{ij} h_{\boldsymbol{\omega}}(\mathbf{x}_i - \mathbf{x}_j) = 2 \sum_{y=1}^{L} (c_y^2 + s_y^2) - \left(\sum_{y=1}^{L} c_y \right)^2 - \left(\sum_{y=1}^{L} s_y \right)^2 - n \,.$$

Therefore,

$$\begin{aligned}
\widehat{\mathcal{L}}_S(h_{\boldsymbol{\omega}}) &= \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \ell(h_{\boldsymbol{\omega}}(\boldsymbol{\delta}_{ij}), \lambda_{ij}) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \frac{1 - \lambda_{ij} h_{\boldsymbol{\omega}}(\boldsymbol{\delta}_{ij})}{2} \\
&= \frac{1}{2} - \frac{1}{2n(n-1)} \sum_{i \neq j}^{n} \lambda_{ij} h_{\boldsymbol{\omega}}(\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{2} - \frac{1}{2n(n-1)} \left[2 \sum_{y=1}^{L} (c_y^2 + s_y^2) - \left(\sum_{y=1}^{L} c_y \right)^2 - \left(\sum_{y=1}^{L} s_y \right)^2 - n \right] \\
&= \frac{n}{2(n-1)} - \frac{1}{2n(n-1)} \left[2 \sum_{y=1}^{L} (c_y^2 + s_y^2) - \left(\sum_{y=1}^{L} c_y \right)^2 - \left(\sum_{y=1}^{L} s_y \right)^2 \right] .
\end{aligned}$$

□

A.3 Experiments

Implementation details. The code used to run the experiments is available at:

https://github.com/gletarte/pbrff

In Section 6 we use the following datasets:

ads http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements
The first 4 features which have missing values are removed.

adult https://archive.ics.uci.edu/ml/datasets/Adult

breast https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

farm https://archive.ics.uci.edu/ml/datasets/Farm+Ads

mnist http://yann.lecun.com/exdb/mnist/
As Sinha and Duchi (2016), binary classification tasks are compiled with the following digits pairs: 1 vs. 7, 4 vs. 9, and 5 vs. 6.

We split the datasets into training and testing sets with a 75/25 ratio except for adult which has a training/test split already computed. We then use 20% of the training set for validation. Table 2 presents an overview. We use the following parameter values range for selection on the validation set:

- $C \in \{10^{-5}, 10^{-4}, \ldots, 10^{4}\}$
- $\sigma \in \{10^{-7}, 10^{-6}, \ldots, 10^{2}\}$
- $\rho \in \{10^{-4}N, 10^{-3}N, \ldots, 10^{0}N\}$
- $\beta \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$
- $D \in \{8, 16, 32, 64, 128\}$

Dataset	n_{train}	n_{valid}	n_{test}	d
ads	1967	492	820	1554
adult	26048	6513	16281	108
breast	340	86	143	30
farm	2485	622	1036	54877
mnist17	9101	2276	3793	784
mnist49	8268	2068	3446	784
mnist56	7912	1979	3298	784

Table 2: Datasets overview.

Supplementary experiments. Figures 4 and 6 present extra results obtained for the *landmarks-based learning* experiments (Subsection 6.1). Figure 5 gives extra results for the *greedy kernel learning* experiment (Subsection 6.2).

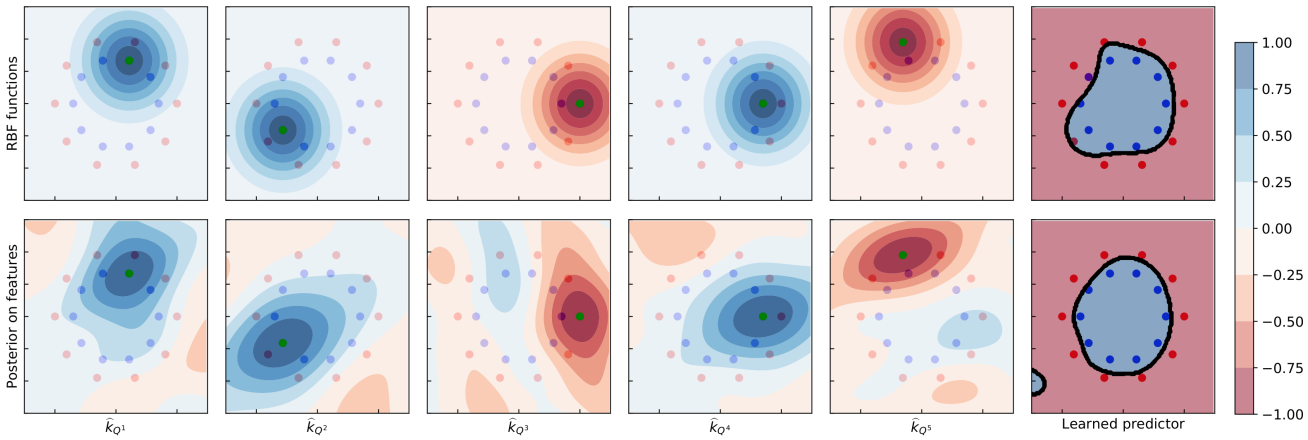
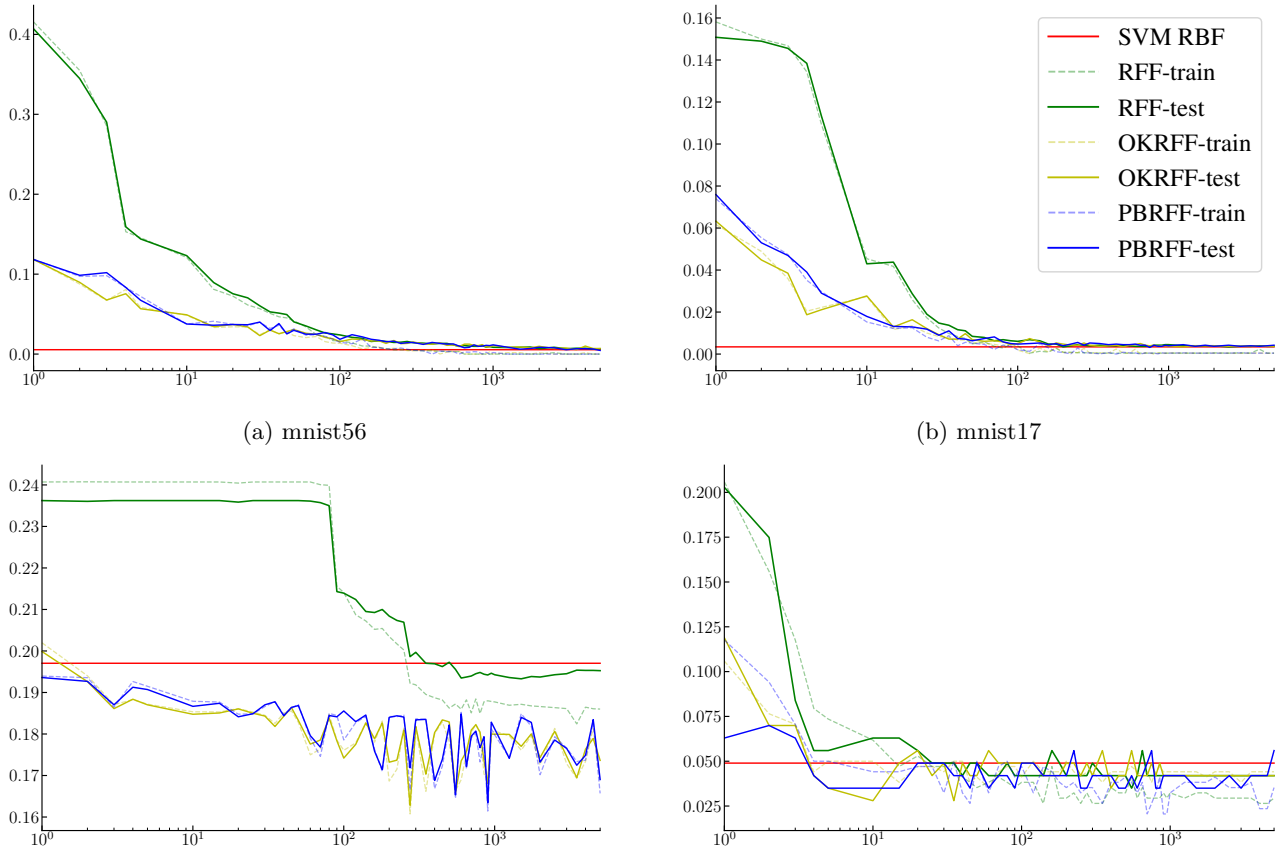


Figure 4: Repetition of Figure 1's experiment, with another toy dataset. First row shows selected RBF-Landmarks kernel outputs, while second row shows the corresponding learned similarity measures on random Fourier features (PB-Landmarks). The rightmost column displays the classification learned by a linear SVM over the mapped dataset.



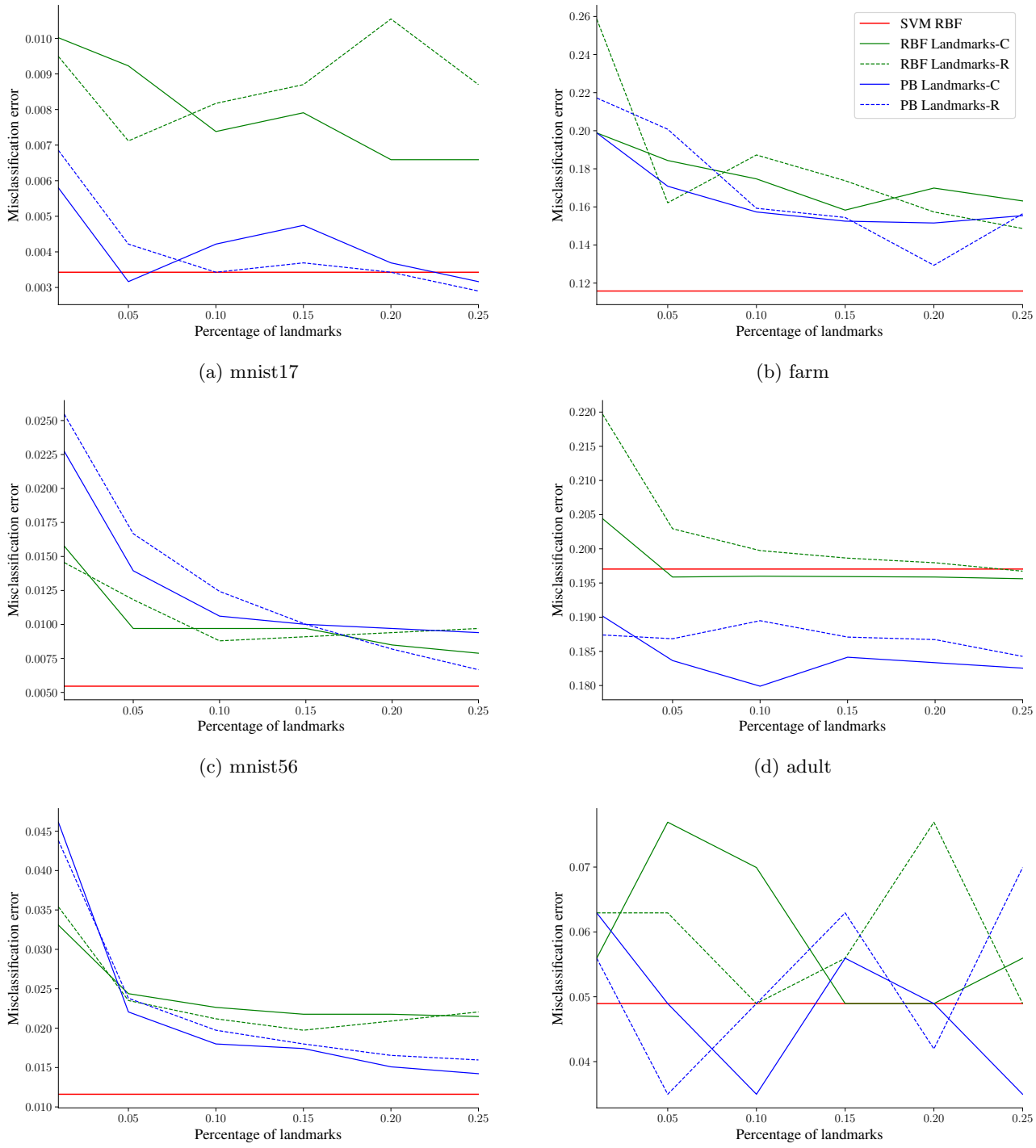
(a) mnist56

(b) mnist17

(c) adult

(d) breast

Figure 5: Train and test error of the kernel learning approaches according to the number of random features D on the remaining 4 datasets (not reported by Figure 3).



(a) mnist17

(b) farm

(c) mnist56

(d) adult

(e) mnist49

(f) breast

Figure 6: Behavior of the landmarks-based approach according to the percentage of training points selected as landmarks on the remaining 6 datasets (not reported by Figure 2).