JMLR: Workshop and Conference Proceedings 44 (2015) 226-241 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

Generalization Bounds for Supervised Dimensionality Reduction

Mehryar Mohri MOHRI@CIMS.NYU.EDU
Courant Institute and Google Research

Afshin Rostamizadeh ROSTAMI@GOOGLE.COM
Google Research

Dmitry Storcheus DSTORCHEUS@GOOGLE.COM
Google Research

Editor: Sanjiv Kumar

Abstract

We introduce and study the learning scenario of *supervised dimensionality reduction*, which couples dimensionality reduction and a subsequent supervised learning step. We present new generalization bounds for this scenario based on a careful analysis of the empirical Rademacher complexity of the relevant hypothesis set. In particular, we show an upper bound on the Rademacher complexity that is in $\widetilde{O}(\sqrt{\Lambda_{(r)}/m})$, where m is the sample size and $\Lambda_{(r)}$ the upper bound on the Ky-Fan r-norm of the operator that defines the dimensionality reduction projection. We give both upper and lower bounds in terms of that Ky-Fan r-norm, which strongly justifies the definition of our hypothesis set. To the best of our knowledge, these are the first learning guarantees for the problem of supervised dimensionality reduction with a *learned* kernel-based mapping. Our analysis and learning guarantees further apply to several special cases, such as that of using a fixed kernel with supervised dimensionality reduction or that of unsupervised learning of a kernel for dimensionality reduction followed by a supervised learning algorithm.

Keywords: PCA, supervised learning, dimensionality reduction, manifold learning, reproducing kernel Hilbert space, learning kernels.

1. Introduction

Dimensionality reduction techniques are common methods in machine learning used either to reduce the computational cost of working in higher-dimensional spaces, or to learn or approximate a manifold expected to be more favorable to a subsequent learning task such as classification or regression. They include classical techniques such as Principal Component Analysis (PCA) (Pearson, 1901) and more recent techniques such as Isometric Feature Mapping (Tenenbaum et al., 2000) and Locally Linear Embedding (Roweis and Saul, 2000). More generally, the dimensionality reduction techniques just mentioned and most others have been shown to be specific instances of the kernel PCA (KPCA) algorithm (Ham et al., 2004), for different choices of a kernel. An even broader view of dimensionality reduction techniques is that they first map input points to the reproducing kernel Hilbert space (RKHS) of some positive semi-definite (PSD) kernel K, and next project vectors onto a low-dimensional space.

Standard dimensionality reduction techniques seek to determine a lower-dimensional space preserving some geometric properties of the input. However, it is not clear which of these properties

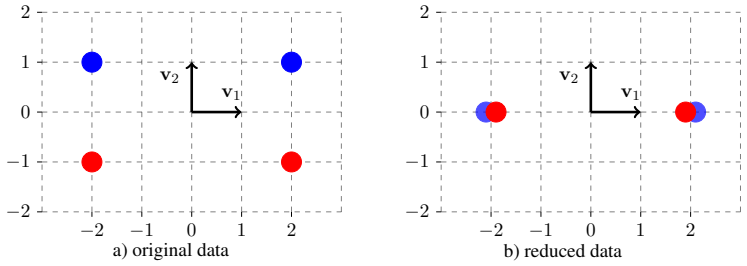

Figure 1: A simple example showing that simply preserving some geometric properties can be detrimental to the subsequent learning task. The original data in (a) has four points from the blue and red classes. The eigenvectors of the covariance matrix are $\mathbf{v}_1 = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $\mathbf{v}_2 = \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$. Standard rank-one PCA projects both blue and red points onto $\mathbf{v}_1$, thus merging them (as plotted in (b)). Any classification on the reduced data will necessarily incur a classification error of at least $\frac{1}{2}$.

would be most beneficial to the later discrimination stage. Since they are typically unsupervised, standard dimensionality reduction techniques also present a risk to the later classification or regression task: the lower-dimensional space found may not be the most helpful one for the second supervised learning stage and, in fact, in some cases could be harmful. Figure 1 shows a very simple example where PCA can lead to a projected space that is detrimental to the subsequent learning stage. More complex variants of this example can occur similarly in higher dimension and for the broader case of KPCA, which covers most known techniques. How should we design dimensionality reduction techniques to most benefit the subsequent supervised learning stage?

This paper seeks precisely to create a theoretical foundation guiding the design of dimensionality reduction with learning guarantees. To do so, we consider a scenario where the dimensionality reduction step is not carried out *blindly* and where, instead, it is coupled with the subsequent supervised learning stage. Since, as already discussed, the key choice defining dimensionality reduction is that of a mapping to an RKHS defined by some PSD kernel K, the learning problem then consists of selecting a PSD kernel K out of a family $\mathcal{K}$ such that the hypothesis learned on a low-dimensional space after projection admits a small generalization error. We call this the *supervised kernel projection* (SKP) setting.

The framework just described bears some similarity with that of *learning kernels* (Lanckriet et al., 2004; Cortes et al., 2009, 2010; Kloft et al., 2011) (see (Gönen and Alpaydın, 2011) and references therein for a recent survey). However, while the selection of a kernel is common to both frameworks, the learning problems and analyses are distinct, in particular because of the learner's freedom to select a projection space after mapping to an RKHS in the dimensionality reduction case. Nevertheless, we will adopt the same common choice for the family $\mathcal{K}$ as in much of the literature for learning kernels, that is that of convex combinations of p base PSD kernels. The RKHS we consider is thus associated to a kernel in that family $\mathcal{K}$ and the projection is over the top r eigenspace of an operator that is a function of the covariance operators of the weighted base kernels. For the scenario of learning kernels, tight generalization bounds are known for this choice of $\mathcal{K}$ (Cortes et al., 2010). The main contribution of this paper is to similarly derive generalization bounds for the SKP framework. Note that, while we consider a broader framework, our generalization bounds also apply to the special case of algorithms proceeding in two decoupled stages of dimensionality reduction followed by supervised learning with a linear model in an RKHS.

The choice of our learning framework is further justified by some previous empirical studies showing that tuning a dimensionality reduction algorithm in a supervised fashion, i.e. taking into account the subsequent learning algorithm using the reduced features, can result in a considerably better performance (Fukumizu et al., 2004; Gönen, 2014). Some recent work also explores learning kernels in the setting of dimensionality reduction (Lin et al., 2011), though no theoretical analysis or justification is provided for the algorithms considered. The vast majority of existing theoretical analyses of dimensionality reduction techniques, even with a fixed kernel, do not directly take into consideration the subsequent learning task and, instead, focus on the optimization of surrogate metrics such as maximizing the variance of the projected features (Zwald and Blanchard, 2005). One exception is the work of Mosci et al. (2007), which provides a generalization guarantee for learning with hypotheses defined by KPCA with a fixed kernel followed by a regression algorithm minimizing the squared loss. Dhillon et al. (2013) also shows that the risk of PCA combined with ordinary least squares regression is at most 4 times that of ridge regression. Recent related work also includes that of Gottlieb et al. (2013), which derives Rademacher complexity generalization bounds for learning Lipschitz functions in a general (fixed) metric space. They show that the intrinsic dimension of the data can significantly influence learning guarantees by bounding the corresponding Rademacher complexity in terms of dimension of underlying manifold and the distortion of training set relative to that manifold.

The results of this paper are organized as follows. In Section 2 and 3 , we describe in detail the learning scenario and the hypothesis set we consider. Section 4 presents our main results, which include an upper bound on the empirical Rademacher complexity of the hypothesis set, and our main generalization bound. In Section 5, we show a lower bound on the sample Rademacher complexity as well as other quantities, which demonstrates a necessary dependence on several crucial quantities and helps to validate the design of the suggested hypothesis class. Finally, in Section 6, we briefly discuss several implications of our results.

2. Learning scenario

Let $\mathcal{X}$ denote the input space. We assume that the learner receives a labeled sample of size m, $S = ((x_1, y_1), \ldots, (x_m, y_m))$, drawn i.i.d. according to some distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$, as well as an unlabeled sample $U = (x'_1, \ldots, x'_u)$ of size u, typically with $u \gg m$, drawn i.i.d. according to the marginal distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$.

We assume that the learner has access to p PSD kernels $K_1, \ldots, K_p$. Instead of requiring the learner to commit to a specific kernel K defining an RKHS, we consider the case of an RKHS defined by a kernel $K_{\boldsymbol{\mu}} = \sum_{k=1} \mu_k K_k$ that is a convex combination of $K_1, \ldots, K_p$. The non-negative mixture weights μ_k, $k = 1, \ldots, p$, are parameters that can be selected by the learner to minimize the error of the classifier using the result of the dimensionality reduction (see Figure 2). The hypothesis set H we consider is thus that of linear hypotheses in a space obtained after projection in the RKHS $\mathbb{H}$ defined by $K_{\boldsymbol{\mu}}$:

$$H = \left\{ x \mapsto \langle w, \Pi_U \Phi(x) \rangle_{\mathbb{H}} \colon \|w\|_{\mathbb{H}} \leq 1, \boldsymbol{\mu} \in \mathcal{M} \right\}. \tag{1}$$

Here, $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\| \cdot \|_{\mathbb{H}}$ denote the inner product and norm in $\mathbb{H}$, $\Phi \colon \mathcal{X} \to \mathbb{H}$ is the feature mapping associated to $K_{\boldsymbol{\mu}}$, Π_U a projection using the unlabeled set U, and $\mathcal{M}$ a regularization set out of which $\boldsymbol{\mu}$ is selected. Note that to avoid a heavier notation, we do not explicitly indicate the dependency of Φ on $\boldsymbol{\mu}$ as this should be clear from the context.

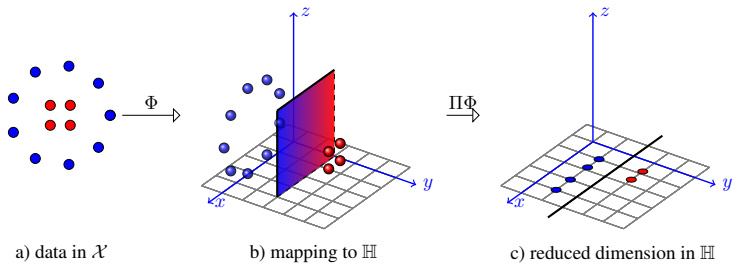

Figure 2: Illustration of the supervised learning scenario: (a) raw input points; (b) points mapped to a higher-dimensional space where linear separation is possible but where not all dimensions are relevant; (c) projection over a lower-dimensional space preserving linear separability.

We now specify the choices of Π_U and $\boldsymbol{\mu}$. For any $k \in [1, p]$, let $C_{U,k} \colon \mathbb{H}_k \to \mathbb{H}_k$ denote the empirical covariance operator based on the unlabeled sample U associated to the PSD kernel $\mu_k K_k$ with RKHS $\mathbb{H}_k$. Let C_U be the operator defined by $C_U = C_{U,1} + \cdots + C_{U,p}$, which acts on the sum of reproducing spaces $\mathbb{H} = \mathbb{H}_1 + \cdots + \mathbb{H}_p$. For a fixed r, Π_U is the rank-r projection over the eigenspace of C_U that corresponds to the top-r eigenvalues of C_U denoted by $\lambda_1(C_U) \geq \ldots \geq \lambda_r(C_U)$.[1] We define similarly the operators $C_{S,k}$ and C_S for the sample S, as well as the projection Π_S.

Note that the dimensionality reduction method just described is in general somewhat different from the standard KPCA with kernel K_μ. Of course, in both cases, the projection is onto the top-r eigenspace of an operator. But, while for KPCA that operator is the empirical covariance operator associated to kernel $K_{\boldsymbol{\mu}}$, in the setting just described, the operator C_U is the sum of the empirical covariance operators associated to each base kernel. In the special case $p = 1$, the two methods coincide. Also, both methods fall into the general SKP framework. Here, we consider the operator C_U or C_S as they admit a favorable structure further discussed in Section 3.

We define the set $\mathcal{M}$ of admissible weight vectors $\boldsymbol{\mu}$ as follows:

$$\mathcal{M} = \left\{ \boldsymbol{\mu} \in \mathbb{R}^p \colon \|\boldsymbol{\mu}\|_{(r)} \leq \Lambda_{(r)}, \;\; \|\boldsymbol{\mu}\|_1 \leq 1, \;\; \sum_{k=1}^{p} \frac{1}{\mu_k} \leq \nu, \;\; \boldsymbol{\mu} \geq 0 \right\}, \tag{2}$$

where $\Lambda_{(r)} \geq 0$ and $\nu \geq 0$ are hyperparameters and where $\|\boldsymbol{\mu}\|_{(r)}$ is the Ky-Fan r-norm of C_U (Bhatia, 1997):[2]

$$\|\boldsymbol{\mu}\|_{(r)} = \|C_U\|_{(r)} = \sum_{i=1}^{r} \lambda_i(C_U)\,. \tag{3}$$

We will later show that this choice of regularization is key as it appears as a crucial term in generalization guarantees and in lower bounds. The vector $\boldsymbol{\mu}$ is further upper bounded by an L_1-norm inequality $\|\boldsymbol{\mu}\|_1 \leq 1$ as is standard in the learning kernel literature with similar kernel combinations. The lower bound constraint on $\boldsymbol{\mu}$, $\sum_{k=1}^{p} \mu_k^{-1} \leq \nu$, implies an upper bound on the eigengap of the induced covariance operator, which is a fundamental quantity that influences the concentration of eigenspaces. In Section 5, we give a simple example demonstrating that the dependency on the eigengap is tight, which implies the necessity of this lower bound regularization.

1. To simplify the presentation, we assume that the selected dimension r satisfies $\lambda_r(C_U) \neq \lambda_{r+1}(C_U)$, but this assumption is not necessary and our results can be straightforwardly extended to more general cases. Note that this assumption is satisfied in particular when the eigenvalues are simple.
2. The Ky-Fan r-norm is in fact a semi-norm.

3. Kernel properties

In this section, we discuss the properties assumed about the base kernels, which turn out to be rather mild assumptions. We will assume that the base kernels K_k, $k \in [1, p]$, satisfy the condition $K_k(x, x) \leq 1$ for all $x \in \mathcal{X}$, which is guaranteed to hold for all normalized kernels, such as Gaussian kernels. We also assume that C_U admits at least r non-zero eigenvalues and that at least one kernel matrix among those associated to kernel K_k on sample S admits rank at least r, and similarly for the kernel matrices defined over the sample U.

We denote by $\mathbf{K}$ the kernel matrix of a kernel K associated to the sample S, $[\mathbf{K}]_{i,j} = K(x_i, x_j)$ and by $\overline{\mathbf{K}}$ the normalized kernel matrix defined by $\overline{\mathbf{K}} = \frac{\mathbf{K}}{m}$. Note that a kernel matrix thereby normalized admits the same eigenvalues as the corresponding sample covariance operator (see for example (Rosasco et al., 2010) Proposition 9.2). In particular, for any $k \in [1, p]$ and $i \in [1, m]$, we have $\lambda_i(\overline{\mathbf{K}}_k) = \lambda_i(C_{S,k})$.

We will assume the base kernels are *linearly independent with respect to the union of the samples S and U*.

Definition 1 (Linearly Independent Kernels) *Let $K_1, \ldots, K_p$ be p PDS kernels and let $S = (x_1, \ldots, x_n)$ be a sample of size m. For any $k \in [1, p]$, let $\mathbb{H}_k$ denote the RKHS associated to K_k and $\overline{\mathbb{H}}_k$ the subspace of $\mathbb{H}_k$ spanned by the set of functions $\{\Phi_{K_k}(x_i) \colon i = 1, \ldots, m\}$. Then, $K_1, \ldots, K_p$ are said to be* linearly independent with respect to the sample S *if, for any $k \in [1, p]$, no non-zero function in $\overline{\mathbb{H}}_k$ can be expressed as a linear combination of the functions in $\cup_{l \neq k} \overline{\mathbb{H}}_l$.*

This condition typically holds in practice, e.g., for polynomial and Gaussian kernels on $\mathbb{R}^N$. As an example, let $\mathcal{X} = \mathbb{R}^N$ and define the sample $S = \{x_1, \ldots, x_m\}$. Define two base kernels: Gaussian $K_1(x, y) = e^{-\|x-y\|^2}$ and linear $K_2(x, y) = \langle x, y \rangle$. Then Φ_{K_1} is defined by $\Phi_{K_1}(x) \colon t \mapsto e^{-\|x-t\|^2}$, that is $\Phi_{K_1}(x)$ is an exponential function $e^{-\|x-t\|^2}$ with parameter x and argument t. Similarly, Φ_{K_2} is defined by $\Phi_{K_2}(x) \colon t \mapsto \langle x, t \rangle$. Thus, $\overline{\mathbb{H}}_1$ is the span of exponential functions $\{e^{-\|x_1-t\|^2}, \ldots, e^{-\|x_m-t\|^2}\}$ and $\overline{\mathbb{H}}_2$ is the span of linear functions $\{\langle x_1, t \rangle, \ldots, \langle x_m, t \rangle\}$. Clearly, no exponential function can be represented as a linear combination of linear functions and likewise, in general, no linear function is represented as a (finite) linear combination of exponential functions. Thus, the base kernels K_1 and K_2 are linearly independent with respect to a finite sample S as in Definition 1. More generally, the support of the base kernels can be straightforwardly modified to ensure that this condition is satisfied.

By definition of $\mathbb{H} = \mathbb{H}_1 + \cdots + \mathbb{H}_p$ and by the results of (Aronszajn, 1950, Section 6), when the base kernels K_k are linearly independent with respect to sample S, then $\overline{\mathbb{H}}_k$ are orthogonal subspaces of $\mathbb{H}$, thus we can define $\overline{\mathbb{H}} = \bigoplus_{k=1}^{p} \overline{\mathbb{H}}_k$, which will be extremely useful in decomposing the spectra of operators C_S. Linearly independent base kernels imply that C_S admits at most pm nonzero eigenvalues of the form $\mu_k \lambda_j(C_{S,k})$.

4. Generalization bound

In this section, we present our generalization bound for learning with the hypothesis set H we introduced in Section 2. To obtain our generalization bound, we derive an upper bound on the empirical Rademacher complexity of H for a sample $S = (x_1, \ldots, x_m)$, which is defined by

$$\widehat{\mathfrak{R}}_S(H) = \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right].$$

Here, σ_is are i.i.d. random variables taking values $+1$ and -1 with equal probabilities. The hypothesis set H we consider is parametrized by w and $\boldsymbol{\mu}$, thus $\widehat{\mathfrak{R}}_S(H)$ can be rewritten as follows:

$$\widehat{\mathfrak{R}}_S(H) = \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\|w\|\leq 1\\ \boldsymbol{\mu}\in\mathcal{M}}}\left\langle w, \Pi_U \sum_{n=1}^{m}\sigma_i\Phi(x_i)\right\rangle\right] = \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\mu}\in\mathcal{M}}\left\|\Pi_U \sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right],$$

where we used the equality case of the Cauchy-Schwarz inequality. To bound the resulting expression, it will be more convenient to work with Π_S instead of Π_U, since we are projecting instances from sample S, and similarly control the Ky-Fan r-norm $\|C_S\|_{(r)}$ rather than $\|C_U\|_{(r)}$. Both of these issues can be addressed by using concentration inequalities to bound the difference of the projections Π_U and Π_S (Zwald and Blanchard, 2005) as well as the difference of the operators C_U and C_S (Shawe-Taylor and Cristianini, 2003). To that end, we first extend the constraint set $\mathcal{M}$ to a larger one $\mathcal{N}$ defined by

$$\mathcal{N} = \left\{\boldsymbol{\mu}\in\mathbb{R}^p\colon \|C_S\|_{(r)} \leq \Lambda_{(r)} + \kappa,\ \ \|\boldsymbol{\mu}\|_1 \leq 1,\ \ \sum_{k=1}^{p}\frac{1}{\mu_k}\leq\nu,\ \ \boldsymbol{\mu}\geq 0\right\}, \tag{4}$$

where $\kappa = 4\Big(1+\sqrt{\frac{\log\left(\frac{2p}{\delta}\right)}{2}}\Big)$. Then, the following lemma provides an upper bound in terms of Π_S.

Lemma 2 *For any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for any $u\in\mathbb{H} = \mathbb{H}_1 + \cdots + \mathbb{H}_p$:*

$$\sup_{\boldsymbol{\mu}\in\mathcal{M}}\|\Pi_U u\| \leq \sup_{\boldsymbol{\mu}\in\mathcal{N}}\left(\|\Pi_S u\| + \frac{8\kappa\nu\|u\|}{\Delta_r\sqrt{m}}\right), \tag{5}$$

where $\Delta_r = \min_{k\in[1,p]}\big(\lambda_r(C_k) - \lambda_{r+1}(C_k)\big)$, C_k is the population covariance operator of kernel K_k and $\kappa = 4\Big(1+\sqrt{\frac{\log\left(\frac{2p}{\delta}\right)}{2}}\Big)$.

The proof of this lemma is given in Appendix A. In view of this lemma, with probability at least $1-\delta$, $\widehat{\mathfrak{R}}_S(H)$ can be bounded as follows

$$\begin{aligned}\widehat{\mathfrak{R}}_S(H) &\leq \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\mu}\in\mathcal{N}}\left(\left\|\Pi_S\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\| + \frac{8\kappa\nu\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\|}{\Delta_r\sqrt{m}}\right)\right]\\ &\leq \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\mu}\in\mathcal{N}}\left\|\Pi_S\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right] + \Big(\frac{8\kappa\nu}{\Delta_r\sqrt{m}}\Big)\frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\mu}\in\mathcal{N}}\left\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right],\end{aligned}$$

using the sub-additivity of the supremum operator and the linearity of expectation. The second term can be bounded as follows:

$$\frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\mu}\in\mathcal{N}}\left\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right] \leq \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\boldsymbol{\mu}\|_1\leq 1}\left\|\sum_{i=1}^{m}\sigma_i\Phi(x_i)\right\|\right] \leq \sqrt{\frac{\eta_0 e\lceil\log p\rceil}{m}}, \tag{6}$$

where $\eta_0 = \frac{23}{22}$, using the bound on the Rademacher complexity of learning kernels given by Theorem 2 of Cortes et al. (2010). The following lemma helps us bound the first term.

Lemma 3 *For any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds:*

$$\frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\sup_{\boldsymbol{\mu} \in \mathcal{N}} \left\| \Pi_S \sum_{i=1}^{m} \sigma_i \Phi(x_i) \right\| \right] \leq \sqrt{\frac{2\big(\Lambda_{(r)} + \kappa\big) \log (2pm)}{m}}, \tag{7}$$

where $\kappa = 4\Big(1 + \sqrt{\frac{\log \frac{2p}{\delta}}{2}}\Big)$.

The proof of the lemma is given in Appendix B. Combining Lemmas 2 and 3 yields directly the following result.

Theorem 4 *Let H be the hypothesis set defined in* (1)*. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. labeled sample S of size $m < u$, the empirical Rademacher complexity of the hypothesis set H can be bounded as follows:*

$$\widehat{\mathfrak{R}}_S(H) \leq \frac{1}{\sqrt{m}} \left[\sqrt{2\big(\Lambda_{(r)} + \kappa\big) \log (2pm)} + \frac{8\kappa\nu}{\Delta_r} \sqrt{\frac{\eta_0 e \lceil \log p \rceil}{m}} \right], \tag{8}$$

where $\Delta_r = \min_{k \in [1,p]} \big(\lambda_r(C_k) - \lambda_{r+1}(C_k)\big)$, C_k is the population covariance operator of kernel K_k, $\kappa = 4\Big(1 + \sqrt{\frac{\log (2p/\delta)}{2}}\Big)$ and $\eta_0 = \frac{23}{22}$.

Note that Δ_r is not a random variable and does not depend on the choice of S or U. It only depends on the spectral properties of the covariance operator for the distribution $\mathcal{D}_{\mathcal{X}}$ and the choice of the projection dimension r.

We now compare this bound to the one known for the Rademacher complexity of a similar hypothesis set in the scenario of learning kernels where a convex combination kernel $K_{\boldsymbol{\mu}}$ is also used (Cortes et al., 2010). This will help us measure the additional complexity cost due to the dimensionality reduction step. Of course, the learning kernel scenario and regularization differ from ours. But, we can make them comparable by considering the case $U = S$, that is the case where U is an unlabeled version of S and can express $\Lambda_{(r)}$ in terms of unscaled sample kernel matrices as follows:

$$\Lambda_{(r)} = \frac{1}{m} \sup_{|I|=r} \sum_{(k,j) \in I} \mu_k \lambda_j(\mathbf{K}_k) \leq \frac{1}{m} \sup_{|I|=r} \sum_{(k,j) \in I} \lambda_j(\mathbf{K}_k). \tag{9}$$

If we define $s_r = \sup_{|I|=r} \sum_{(k,j) \in I} \lambda_j(\mathbf{K}_k)$ as the largest r-sum of eigenvalues selected from all base kernel matrices, and $s'_m = \sup_{k \in [1,p]} \operatorname{Tr}[\mathbf{K}_k]$, which is the largest m-sum of eigenvalues selected from a single base kernel matrix, the Rademacher complexity of our hypothesis class is in $\widetilde{O}\big(\sqrt{s_r/m}\big)$, while that of the hypothesis used in the learning kernel setting is in $\widetilde{O}\big(\sqrt{s'_m/m}\big)$. Thus, for $r = m$, the upper bound on the Rademacher complexity in our supervised dimensionality case is higher. The difference is due precisely to the extra freedom that the learner has to define a projection space by selecting eigenvectors from different kernel matrices, while in the learning kernel case he needs to commit instead to a single kernel matrix. For r sufficiently smaller than m, the complexity term in the supervised dimensionality case could of course be more favorable ($s_r \leq s'_m$).

The following is our main generalization bound for supervised dimensionality reduction. We denote by $R(h)$ the generalization error with respect to the zero-one loss and by $\widehat{R}_{S,\rho}(h)$ the empirical margin loss of $h \in H$, that is the fraction of points in S classified with margin less than ρ by h.

Theorem 5 *Let H be the hypothesis set defined in* (1)*. Then, with probability at least $1-\delta$ over the draw of a sample S of size m, the following holds for all $h \in H$:*

$$R(h) \leq \widehat{R}_{S,\rho}(h) + \frac{2}{\rho\sqrt{m}}\left(\sqrt{2\big(\Lambda_{(r)}+\kappa\big)\log\left(2pm\right)} + \frac{8\kappa\nu}{\Delta_r}\sqrt{\frac{\eta_0 e\lceil\log p\rceil}{m}}\right) + 3\sqrt{\frac{\log\frac{4p}{\delta}}{2m}}\,.$$

Proof The theorem follows directly by combining the high-probability upper bound on the Rademacher complexity given by Theorem 4 and the standard high-probability Rademacher-based generalization bound of Koltchinskii and Panchenko (2002) (see also (Bartlett and Mendelson, 2003)). ■

To our knowledge, this is the first learning guarantee given for the scenario of supervised dimensionality reduction. The bound of the theorem is in $O\big(\sqrt{\Lambda_{(r)}\log(pm)/m}\big)$. Thus, it suggests that the Ky-Fan r-norm of the covariance operator plays a key role in the generalization ability of hypotheses in this scenario. This is further supported by the presence of that term in a lower bound proven in the next section. Note that the dependency of the bound on the number of base kernels p is only logarithmic, which suggests using a rather large number of base kernels. The presence of the term in the bound depending on ν and Δ_r is due to the concentration bound for projections. The parameter ν controls the eigengap of the learned operator C_U, while, as already pointed out, Δ_r is a quantity that does not depend on the sample or on $\boldsymbol{\mu}$, it is entirely defined by the choice of the base kernel functions. We further elaborate on this in Section 5.

We note that Mosci et al. (2007) and Gottlieb et al. (2013) also give generalization bounds for a supervised scenario of dimensionality reduction, however, they do not learn a mapping and projection for dimensionality reduction jointly with a hypothesis learned on the projected space using a discriminative algorithm. Nevertheless, their generalization bounds are comparable to the special case of our bound where $p = 1$.

The analysis of Gottlieb et al. (2013) is presented for general metric spaces, which is more general than what we consider here. In the case of the Euclidean space, their generalization bound is in $O\big(\sqrt{d/m}+\sqrt{\eta/m}\big)$, where d is the dimension of underlying data manifold and η is the average distance of the training set to that manifold. While both bounds admit a similar dependence on m, our bound relies on the Ky-Fan norm of the projection rather than the intrinsic dimension of the dataset. We note that the existence of an approximate low-dimensional manifold is a distributional assumption which, depending on the task, may not hold. Furthermore, even when it does, the estimation of the intrinsic dimension is typically a difficult task. The Ky-Fan norm, on the other hand, can be directly controlled by the choice of the regularization parameter in the definition of the hypothesis set.

The generalization bound of Mosci et al. (2007) is in $O\big(1/\sqrt{m}\big)$. However, while we fix the number of eigenvalues for dimensionality reduction to r, their bound requires selecting all eigenvalues above a threshold $\lambda_m = O\big(1/\sqrt{m}\big)$. Furthermore, as already mentioned, their analysis holds for the specific setting of KPCA with a fixed kernel ($p = 1$) followed by Ridge regression.

5. Lower bounds

In this section, we show a lower bound on the Rademacher complexity of the hypothesis class H defined by (1). Furthermore, we give a simple example demonstrating the necessity of the eigengap term appearing in Lemma 2 and also motivate the additional regularization term ν.

Theorem 6 *For any m and r there exist samples S and U, a setting of the regularization parameter $\Lambda_{(r)}$, as well as a choice of base kernels $K_1, \ldots, K_p$ such that the following inequality holds:*

$$\widehat{\mathfrak{R}}_S(H) \geq \sqrt{\frac{\Lambda_{(r)}}{2m}}\,.$$

The proof is given in Appendix C. The result proves the tightness of the upper bound we derived in terms of m, up to logarithmic factors. It further shows the key role of the regularization parameter $\Lambda_{(r)}$ and justifies the presence of Ky-Fan r-norm constraint in the definition of the hypothesis set.

We now also give a simple example showing that the projections must necessary depend on an eigengap quantity. This in turn motivates the dependency of Lemma 2 on the quantity Δ_r as well as the regularization $\sum_{k=1}^{p} \mu_k^{-1} \leq \nu$ which is used to bound the eigengap of the learned operator C_S (see equation (12) in the proof of Lemma 2). The fact that the eigengap is essential for the concentration of projections has been known in the matrix perturbation theory literature (Stewart and Sun, 1990). The following proposition gives an example which shows that the dependence on the eigengap is tight.

Proposition 7 *There exist operators A and B such that*

$$\|P_r(A) - P_r(B)\| = \frac{2\|A - B\|}{\lambda_r(A) - \lambda_{r+1}(A)},$$

where $P_r(A)$ (resp. $P_r(B)$) is the orthogonal projection onto the top r eigenspace of A (resp. B).

Proof Let $r = 1$ and consider A and B defined as follows: $A = \left(\begin{smallmatrix} 1+\epsilon & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1+\epsilon \end{smallmatrix}\right)$, thus $A - B = \left(\begin{smallmatrix} \epsilon & 0 \\ 0 & -\epsilon \end{smallmatrix}\right)$, which implies that $\|A - B\| = \epsilon$. Also, the eigengap is equal to $\lambda_1(A) - \lambda_2(A) = \epsilon$. Now, note that $P_1(A)$ is the projection onto $e_1 = (1, 0)^\top$ and $P_1(B)$ is the projection onto $e_2 = (0, 1)^\top$. Since e_1 and e_2 are orthogonal, this implies $\|P_1(A) - P_1(B)\| = \|P_1(A)\| + \|P_1(B)\| = 2$. On the other hand, $\frac{2\|A-B\|}{\lambda_1(A)-\lambda_2(A)} = \frac{2\epsilon}{\epsilon} = 2$, which completes the proof. ■

6. Discussion

We now briefly discuss the results presented in the previous sections. Let us first emphasize that our choice of the hypothesis class H (Section 2) is strongly justified a posteriori by the learning guarantees we presented: both our upper and lower bounds on the Rademacher complexity (Sections 4 and 5) suggest that the quantities present in the definition of H play an important role. The regularization parameters we provide can be tuned to directly bound each of these crucial quantities and thereby limit the risk of over fitting.

Second, the hypothesis set suggested in this paper provides a unified framework for choosing an optimal dimensionality reduction method. It suggests to specify a set of potential methods (equivalent to a set of base kernels) and then learn their combination jointly with a projection. Moreover,

the generalization bound is logarithmic in the number of base kernels, which encourages the use of a very large base set.

Third, we observe that the hypothesis class H clearly motivates the design of a single-stage coupled algorithm. Such an algorithm would be based on structural risk minimization (SRM) and seek to minimize the empirical error over increasingly complex hypothesis sets, by varying the parameters $\Lambda_{(r)}$ and ν, to trade-off empirical error and model complexity. It is worthwhile to note that our hypothesis set is constructed in such a way that the search over the choices of parameters $\boldsymbol{\mu}$ does not incur the bottleneck of recomputing the eigendecomposition of operator C_U at every iteration. Instead, we require the computation of the eigendecomposition of the (unweighted) base kernel matrices once as a preprocessing step. The key to that is the assumption of linearly independent kernels, which is typically satisfied in practice.

We note that the existing literature has empirically evaluated both learning kernels with KPCA in an unsupervised (two-stage) fashion (Zhuang et al., 2011; Lin et al., 2011) and applied supervised KPCA (single-stage training) with a fixed kernel function (Fukumizu et al., 2004; Gönen, 2014). While these existing algorithms do not directly consider the hypothesis class we motivated, they can, in certain cases, still select a hypothesis function that is found in our class. In particular, our learning guarantees are applicable to hypotheses chosen in a two-stage manner, as long as the regularization constraints are satisfied and the same family of projections are used. Similarly, the case $p = 1$, which corresponds to the standard fixed-kernel supervised learning scenario, is covered by our analysis. Even in such cases, the bounds that we provide would be the first to guarantee the generalization ability of the algorithm via bounding the sample Rademacher complexity.

7. Conclusion

We presented a new analysis and generalization guarantees for the scenario of supervised dimensionality reduction with a learned kernel. The hypothesis class is designed with regularization constraints that are directly motivated by the upper and lower bounds on its Rademacher complexity. Our analysis suggests the design of learning algorithms for selecting hypotheses from this specifically tailored class, either in a two-stage or a single-stage manner. Our analysis can also benefit the study of other similar hypothesis sets within the SKP framework.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, pages 337–404, 1950.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

Gilles Blanchard and Laurent Zwald. Finite-dimensional projection for classification and statistical learning. *Information Theory, IEEE Transactions on*, 54(9):4169–4182, 2008.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L_2 regularization for learning kernels. In *Proceedings of UAI*, pages 109–116, 2009.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of ICML*, pages 247–254, 2010.

Paramveer S Dhillon, Dean P Foster, Sham M Kakade, and Lyle H Ungar. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14(1): 1505–1511, 2013.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, 5: 73–99, 2004.

Mehmet Gönen. Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning. *Pattern recognition letters*, 38:132–141, 2014.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction. In *Proceedings of ALT*, pages 279–293. Springer, 2013.

Jihun Ham, Daniel D Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of ICML*, page 47. ACM, 2004.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50, 2002.

Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh. Multiple kernel learning for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(6):1147–1160, 2011.

Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, 9(2):245–303, 2000.

Sofia Mosci, Lorenzo Rosasco, and Alessandro Verri. Dimensionality reduction and generalization. In *Proceedings of ICML*, pages 657–664. ACM, 2007.

Fedor L Nazarov and Anatoliy N Podkorytov. Ball, haagerup, and distribution functions. In *Complex analysis, operators, and related topics*, pages 247–267. Springer, 2000.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *The Journal of Machine Learning Research*, 11:905–934, 2010.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

John Shawe-Taylor and Nello Cristianini. Estimating the moments of a random vector with applications. In *Proceedings of GRETSI*, pages 47–52, 2003.

GW Stewart and J Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Jinfeng Zhuang, Jialei Wang, Chu Hong Hoi, and Xiangyang Lan. Unsupervised multiple kernel learning. *Journal of Machine Learning Research (JMLR)*, 20:129–144, 2011.

Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Proceedings of NIPS*, pages 1649–1656, 2005.

Appendix A. Proof of Lemma 2

Proof For the first part of the proof, let $E_k = \mathrm{E}\left[C_{S,k}\right] = \mathrm{E}\left[C_{U,k}\right]$ be the population covariance operator of kernel $\mu_k K_k$. We will show a concentration bound on $C_{S,k}$ and $C_{U,k}$ that holds uniformly over $k \in [1,p]$. Using the union bound, Lemma 1 from (Zwald and Blanchard, 2005) (equivalently Corollary 5 from (Shawe-Taylor and Cristianini, 2003)), and assuming $u > m$, with probability at least $1-\delta$, the following holds for all $k \in [1,p]$,

$$\max\left[\|E_k - C_{S,k}\|_{\overline{\mathbb{H}}_k}, \|E_k - C_{U,k}\|_{\overline{\mathbb{H}}_k}\right] \leq \frac{\mu_k \kappa}{2\sqrt{m}}\,, \tag{10}$$

where $\kappa = 4\Big(1 + \sqrt{\frac{\log\left(\frac{2p}{\delta}\right)}{2}}\Big)$.

Let Π_k be the orthogonal projection onto the top r eigenfunctions of E_k. By decomposing over orthogonal subspaces of $\overline{\mathbb{H}} = \bigoplus_{k=1}^{p} \overline{\mathbb{H}}_k$ as well as adding and subtracting Π_k, we can bound $\|\Pi_S - \Pi_U\|$ by $\sum_{k=1}^{p} \|\Pi_k - \Pi_S\|_{\overline{\mathbb{H}}_k} + \sum_{k=1}^{p} \|\Pi_k - \Pi_U\|_{\overline{\mathbb{H}}_k}$. Now, since $C_{S,k}$ is the restriction of C_S to $\overline{\mathbb{H}}_k$, the following inequality holds for all $k \in [1,p]$:

$$\|\Pi_k - \Pi_S\|_{\overline{\mathbb{H}}_k} \leq \frac{8\|E_k - C_S\|_{\overline{\mathbb{H}}_k}}{\lambda_r(E_k) - \lambda_{r+1}(E_k)} = \frac{8\|E_k - C_{S,k}\|_{\overline{\mathbb{H}}_k}}{\lambda_r(E_k) - \lambda_{r+1}(E_k)}\,. \tag{11}$$

A similar statement holds for the projection with respect to sample U. To obtain the bound above we consider two cases, either $8\|E_k - C_S\|_{\overline{\mathbb{H}}_k}/(\lambda_r(E_k) - \lambda_{r+1}(E_k)) \leq 1/4$, which is a sufficient condition for Theorem 3 of (Zwald and Blanchard, 2005) that directly implies (11). Otherwise, if the condition does not hold, then the right-hand side of (11) will be larger than 2, which is a trivial bound on the difference of two projections.

Next, we use the constraint $\|\boldsymbol{\mu}\|_1 \leq 1$ to upper bound (10) by $\kappa/(2\sqrt{m})$ and lower bound $\lambda_r(E_k) - \lambda_{r+1}(E_k) = \mu_k\big(\lambda_r(C_k) - \lambda_{r+1}(C_k)\big) \geq \mu_k \Delta_r$, where C_k is the population covariance operator of kernel K_k and $\Delta_r = \min_{k\in[1,p]}\big(\lambda_r(C_k) - \lambda_{r+1}(C_k)\big)$. Now $4\kappa/(\sqrt{m}\mu_k\Delta_r)$ is the uniform bound on the norm of projections in (11). Summing up $\|\Pi_k - \Pi_S\|_{\overline{\mathbb{H}}_k} + \|\Pi_k - \Pi_U\|_{\overline{\mathbb{H}}_k}$ over k and applying the uniform bound $4\kappa/(\sqrt{m}\mu_k\Delta_r)$, which holds for both samples U and S, we conclude that the following holds:

$$\|\Pi_S - \Pi_U\| \leq \sum_{k=1}^{p} \frac{1}{\mu_k}\frac{8\kappa}{\Delta_r\sqrt{m}} \leq \frac{8\kappa\nu}{\Delta_r\sqrt{m}}\,. \tag{12}$$

For the second portion of the proof, we use a series of inequalities to show

$$\big|\|C_U\|_{(r)} - \|C_S\|_{(r)}\big| \leq \sum_{i=1}^{r} |\lambda_i(C_U) - \lambda_i(C_S)| \leq \sqrt{r}\Big(\sum_{i=1}^{r} |\lambda_i(C_U) - \lambda_i(C_S)|^2\Big)^{1/2}\,,$$

which is in turn bounded by $\sqrt{r}\|C_U - C_S\|$ using the Hoffman-Wielandt inequality. Next, by decomposing over orthogonal subspaces of $\overline{\mathbb{H}} = \bigoplus_{k=1}^{p} \overline{\mathbb{H}}_k$ together with adding and subtracting E_k, we bound $\|C_U - C_S\|$ by $\sum_{k=1}^{p} \|E_k - C_{S,k}\|_{\overline{\mathbb{H}}_k} + \sum_{k=1}^{p} \|E_k - C_{U,k}\|_{\overline{\mathbb{H}}_k}$. If we again apply the uniform bound from (10) in the form $\mu_k\kappa/2\sqrt{m}$, we obtain that with probability at least $1-\delta$, the following holds:

$$\big|\|C_U\|_{(r)} - \|C_S\|_{(r)}\big| \leq \sum_{k=1}^{p} \frac{\sqrt{r}\mu_k\kappa}{\sqrt{m}} \leq \kappa\,. \tag{13}$$

Combining (12) and (13) yields

$$\sup_{\boldsymbol{\mu}\in\mathcal{M}} \|\Pi_U u\| \leq \sup_{\boldsymbol{\mu}\in\mathcal{N}} \|\Pi_U u\| \leq \sup_{\boldsymbol{\mu}\in\mathcal{N}} \left(\|\Pi_S u\| + \frac{8\kappa\nu}{\Delta_r\sqrt{m}}\|u\| \right), \tag{14}$$

with probability $1-\delta$. ■

Appendix B. Proof of Lemma 3

We will use the following lemma to give the proof of Lemma 3.

Lemma 8 *For each $k \in [1,p]$ let $\lambda_1(\overline{\mathbf{K}}_k) \geq \cdots \geq \lambda_m(\overline{\mathbf{K}}_k)$ be the eigenvalues of $\overline{\mathbf{K}}_k$ with corresponding orthonormal eigenvectors $\mathbf{v}_{k,1}, \cdots, \mathbf{v}_{k,m}$. For $\mu \in \mathcal{N}$, let I_μ denote the set of indices (k,j) corresponding to the largest r elements of the set $\{\mu_k\lambda_j(\overline{\mathbf{K}}_k)\}_{k,j}$, then the following equality holds:*

$$\left\| \Pi_S \sum_{i=1}^m \sigma_i \Phi(x_i) \right\| = \sqrt{m \sum_{(k,j)\in I_{\boldsymbol{\mu}}} \mu_k\lambda_j(\overline{\mathbf{K}}_k)(\mathbf{v}_{k,j}^\top \boldsymbol{\sigma})^2}\,, \tag{15}$$

where $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_m)^\top$.

Proof Recall from Section 3 that when $\overline{\mathbb{H}} = \bigoplus_{k=1}^p \overline{\mathbb{H}}_k$, the eigenvalues of C_S take the form $\mu_k\lambda_j(\overline{\mathbf{K}}_k)$ with orthonormal eigenfunctions $u_{k,j}$, where, for each $k \in [1,p]$, functions $u_{k,1}, \ldots, u_{k,m}$ belong to the orthogonal component $\overline{\mathbb{H}}_k$. Thus, we can write $\|\Pi_S f\|^2 = \sum_{(k,j)\in I_\mu} \langle u_{k,j}, f\rangle^2$, for any $f \in \overline{\mathbb{H}}$. When $f = \Phi(x_i)$, it suffices to take the inner product of $u_{k,j}$ and f in $\overline{\mathbb{H}}_k$, which, by the reproducing property, is equal to $u_{k,j}(x_i)$. By (Blanchard and Zwald, 2008, Equation (8)), $u_{k,j}(x_i)$ takes the form

$$u_{k,j}(x_i) = \sqrt{\frac{\mu_k}{\lambda_j(\overline{\mathbf{K}}_k)m}} \sum_{n=1}^m K_k(x_i,x_n)[\mathbf{v}_{k,j}]_n\,. \tag{16}$$

Since $\mathbf{v}_{k,j}$ is an eigenvector of $\overline{\mathbf{K}}_k$, we see that

$$\sum_{n=1}^m K_k(x_i,x_n)[\mathbf{v}_{k,j}]_n = m\sum_{n=1}^m [\overline{\mathbf{K}}_k]_{i,n}[\mathbf{v}_{k,j}]_n = m\lambda_j(\overline{\mathbf{K}}_k)[\mathbf{v}_{k,j}]_i\,, \tag{17}$$

which means that $u_{k,j}(x_i) = \sqrt{m\mu_k\lambda_j(\overline{\mathbf{K}}_k)}[\mathbf{v}_{k,j}]_i$. In view of this expression, we can write

$$\langle u_{k,j}, \sum_{i=1}^m \sigma_i\Phi(x_i)\rangle = \sum_{i=1}^m \sigma_i u_{k,j}(x_i) = \sqrt{m\mu_k\lambda_j(\overline{\mathbf{K}}_k)}\mathbf{v}_{k,j}^\top\boldsymbol{\sigma}. \tag{18}$$

Squaring the terms above and summing them over the set of indices I_μ completes the proof. ■

The following gives the proof of Lemma 3.

Proof [Lemma 3] The term $\|\Pi_S \sum_{i=1}^m \sigma_i\Phi(x_i)\|$ can be directly bounded using only the Ky-Fan norm constraint on $\|C_S\|_{(r)}$, since it controls the spectrum of the projection. Thus, we will simplify the problem to analyze the supremum over choices of $\boldsymbol{\mu}$ that satisfy $\|C_S\|_{(r)} \leq \epsilon$, where $\epsilon = \Lambda_{(r)} + \kappa$. This clearly includes all elements in $\mathcal{N}$ as well.

For this proof, we will use the representation of $\|\Pi_S \sum_{i=1}^m \sigma_i \Phi(x_i)\|$ from Lemma 8, which will be upper bounded by the supremum over the choice of all size-r sets in order to remove the dependence of the set of indices on the identity of the top eigenvalues:

$$\sup_{\|C_S\|_{(r)} \leq \epsilon} \sum_{(k,j) \in I_{\boldsymbol{\mu}}} \mu_k \lambda_j(\overline{\mathbf{K}}_k)(\mathbf{v}_{k,j}^\top \boldsymbol{\sigma})^2 \leq \sup_{|I|=r} \sup_{\|C_S\|_{(r)} \leq \epsilon} \sum_{(k,j) \in I} \mu_k \lambda_j(\overline{\mathbf{K}}_k)(\mathbf{v}_{k,j}^\top \boldsymbol{\sigma})^2 \,, \tag{19}$$

where $\sup_{|I|=r}$ indicates the supremum over all indexing sets. We can express the sum above as an inner product $\mathbf{u}_\mu \cdot \mathbf{u}_\sigma$, where $\mathbf{u}_\mu$ is an $r-$dimensional vector with entries $\mu_k \lambda_j(\overline{\mathbf{K}}_k)$ and $\mathbf{u}_\sigma$ has entries $(\mathbf{v}_{k,j}^\top \boldsymbol{\sigma})^2$ such that $(k,j) \in I$. By construction, we have $\|C_S\|_{(r)} = \|\mathbf{u}_\mu\|_1$, thus, we will reduce the problem to that of analyzing $\sup_{\|\mathbf{u}_\mu\|_1 \leq \epsilon}$. Then, by definition of the dual norm, we can write:

$$\sup_{|I|=r} \sup_{\|\mathbf{u}_\mu\|_1 \leq \epsilon} \mathbf{u}_\mu \cdot \mathbf{u}_\sigma = \sup_{|I|=r} \epsilon \|\mathbf{u}_\sigma\|_\infty = \epsilon \max_{k,j} (\mathbf{v}_{k,j}^\top \boldsymbol{\sigma})^2 \,. \tag{20}$$

Thus, $\|\Pi_S \sum_{i=1}^m \sigma_i \Phi(x_i)\|$ is bounded by the following:

$$\max_{k,j} \sqrt{m\epsilon (\mathbf{v}_{k,j}^\top \boldsymbol{\sigma})^2} = \sqrt{m\epsilon} \max_{k,j} |\mathbf{v}_{k,j}^\top \boldsymbol{\sigma}| = \sqrt{m\epsilon} \max_{k,j} \max_{s_t \in \{-1,1\}} s_t \mathbf{v}_{k,j}^\top \boldsymbol{\sigma} \,.$$

By Massart's lemma (Massart, 2000), we can write

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\max_{k,j} \max_{s_t \in \{-1,1\}} s_t \mathbf{v}_{k,j}^\top \boldsymbol{\sigma} \right] \leq \sqrt{2 \log (2pm)} \,. \tag{21}$$

This follows since the norm of $s_t \mathbf{v}_{k,j}$ is bounded by 1 and since the cardinality of the set over which the maximum is taken is bounded by $2pm$. Combining all the intermediate results leads to the following:

$$\frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|C_S\|_{(r)} \leq \epsilon} \|\Pi_S \sum_{i=1}^m \sigma_i \Phi(x_i)\| \right] \leq \sqrt{\frac{2\epsilon \log (2pm)}{m}}. \tag{22}$$

The final result is obtained by setting $\epsilon = \Lambda_{(r)} + \kappa$. ■

Appendix C. Proof of Theorem 6

Proof First we let S and U be any two samples, both of size m, such that U is simply an unlabeled version of S. Now, assume that $p = 1$ and that the sample kernel matrix $\overline{\mathbf{K}}_1$ of kernel K_1 admits exactly r distinct non-zero simple eigenvalues. Finally, select $\Lambda_{(r)}$ such that $\Lambda_{(r)} / \lambda_1(\overline{\mathbf{K}}_1) \leq 1$.

As calculated in Section 4, $\sup_{\|w\| \leq 1} \sum_{i=1}^m \sigma_i h(x_i) = \|\Pi_U \sum_{i=1}^m \sigma_i \Phi(x_i)\|$ and in this particular scenario $\|C_U\|_{(r)} = \|C_S\|_{(r)}$. Thus, the empirical Rademacher complexity simplifies to $\widehat{\mathfrak{R}}_S(H) = \frac{1}{m} \mathrm{E}_{\boldsymbol{\sigma}} \left[\sup_{\|C_S\|_{(r)} \leq \Lambda_{(r)}} \|\Pi_S \sum_{i=1}^m \sigma_i \Phi(x_i)\| \right]$, where the projection can be written directly in terms of the sample S. Here, the L_1 constraint on $\boldsymbol{\mu}$ is not needed, since it is satisfied by the Ky-Fan r-norm constraint when $\Lambda_{(r)} \leq \lambda_1(\overline{\mathbf{K}}_1)$.

Now, following the steps from Lemma 8, we can express the norm of the projection as follows:

$$\left\| \Pi_S \sum_{i=1}^m \sigma_i \Phi(x_i) \right\| = \sqrt{m \sum_{j=1}^r \mu_1 \lambda_j(\overline{\mathbf{K}}_1)(\boldsymbol{\sigma}^\top \mathbf{v}_{1,j})^2} \,. \tag{23}$$

Note that here, unlike the general statement of Lemma 8, the choice of the r entries that appear in the sum is not effected by the value of $\boldsymbol{\mu}$, since there are in fact only r non-zero eigenvalues in total, by construction (i.e. there is one base kernel of rank r). The choice of $\boldsymbol{\mu}$, however, still affects the scale of the r eigenvalues.

The expression is furthermore simplified by introducing the vectors $\mathbf{u}_\mu$ with entries $\mu_1 \lambda_j(\overline{\mathbf{K}}_1)$ and $\mathbf{u}_\sigma$ with entries $(\mathbf{v}_{1,j}^\top \boldsymbol{\sigma})^2$, which is similar to the proof of Lemma 3. By the monotonicity of the square-root function and using the definition of $\mathbf{u}_{\boldsymbol{\mu}}$ as well as the dual norm we have

$$\sup_{\|C_S\|_{(r)} \leq \Lambda_{(r)}} \sqrt{\mathbf{u}_\mu \cdot \mathbf{u}_{\boldsymbol{\sigma}}} = \sqrt{\sup_{\|\mathbf{u}_{\boldsymbol{\mu}}\|_1 \leq \Lambda_{(r)}} \mathbf{u}_\mu \cdot \mathbf{u}_{\boldsymbol{\sigma}}} = \sqrt{\Lambda_{(r)} \|\mathbf{u}_\sigma\|_\infty}\,. \tag{24}$$

Thus, the Rademacher complexity is reduced to

$$\hat{\mathfrak{R}}_S(H) = \sqrt{\frac{\Lambda_{(r)}}{m}} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\sqrt{\max_{j \in [1,r]} (\mathbf{v}_{1,j}^\top \boldsymbol{\sigma})^2} \right] = \sqrt{\frac{\Lambda_{(r)}}{m}} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\max_{j \in [1,r]} |\mathbf{v}_{1,j}^\top \boldsymbol{\sigma}| \right]. \tag{25}$$

Finally, we use Jensen's inequality and Khintchine's inequality to show

$$\mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[\max_{j \in [1,r]} |\mathbf{v}_{1,j}^\top \boldsymbol{\sigma}| \right] \geq \max_{j \in [1,r]} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \left[|\mathbf{v}_{1,j}^\top \boldsymbol{\sigma}| \right] \geq \max_{j \in [1,r]} 2^{-1/2} \|\mathbf{v}_{1,j}\| = 2^{-1/2}, \tag{26}$$

where the tight constant $2^{-1/2}$ used in Khintchine's inequality can be found in (Nazarov and Podkorytov, 2000)[Chapter II]. Plugging this constant back into equation (25) completes the proof of the theorem. ■