Electronic Journal of Statistics
Vol. 12 (2018) 2141–2192
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1399

Fast learning rate of non-sparse multiple kernel learning and optimal regularization strategies

Taiji Suzuki

The University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656, Japan.
RIKEN Center for Advanced Intelligence Project
1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan.
e-mail: taiji@mist.i.u-tokyo.ac.jp

Abstract: In this paper, we give a new generalization error bound of Multiple Kernel Learning (MKL) for a general class of regularizations, and discuss what kind of regularization gives a favorable predictive accuracy. Our main target in this paper is dense type regularizations including ℓ_p-MKL. According to the numerical experiments, it is known that the sparse regularization does not necessarily show a good performance compared with dense type regularizations. Motivated by this fact, this paper gives a general theoretical tool to derive fast learning rates of MKL that is applicable to arbitrary mixed-norm-type regularizations in a unifying manner. This enables us to compare the generalization performances of various types of regularizations. As a consequence, we observe that the homogeneity of the complexities of candidate reproducing kernel Hilbert spaces (RKHSs) affects which regularization strategy (ℓ_1 or dense) is preferred. In fact, in homogeneous complexity settings where the complexities of all RKHSs are evenly same, ℓ_1-regularization is optimal among all isotropic norms. On the other hand, in inhomogeneous complexity settings, dense type regularizations can show better learning rate than sparse ℓ_1-regularization. We also show that our learning rate achieves the minimax lower bound in homogeneous complexity settings.

Keywords and phrases: Multiple kernel learning, fast learning rate, minimax lower bound, regularization, generalization error bounds.

Received March 2017.

1. Introduction

Multiple Kernel Learning (MKL) proposed by Lanckriet et al. (2004) is one of the most promising methods that adaptively select the kernel function in supervised kernel learning. Kernel method has been widely used in machine learning and data analysis and several studies have supported its usefulness (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). However the performance of kernel methods critically relies on the choice of the kernel function. Many methods have been proposed to deal with the issue of kernel selection.

Ong et al. (2005) studied hyperkernels as a kernel of kernel functions. Argyriou et al. (2006) considered DC programming approach to learn a mixture of kernels with continuous parameters. Some studies tackled a problem to learn non-linear combination of kernels as in Bach (2009); Cortes et al. (2009a); Varma and Babu (2009). Among them, learning a linear combination of finite candidate kernels with non-negative coefficients is the basic, fundamental and commonly used approach. The seminal work of MKL by Lanckriet et al. (2004) considered learning convex combination of candidate kernels as well as its linear combination. This work opened up the sequence of the MKL studies. Bach et al. (2004) showed that MKL can be reformulated as a kernel version of the group lasso (Yuan and Lin, 2006). This formulation gives an insight that MKL can be described as a ℓ_1-mixed-norm regularized method. As a generalization of MKL, ℓ_p-MKL that imposes ℓ_p-mixed-norm regularization has been proposed (Micchelli and Pontil, 2005; Kloft et al., 2009). ℓ_p-MKL includes the original MKL as a special case as ℓ_1-MKL. Another direction of generalization is elasticnet-MKL (Shawe-Taylor, 2008; Tomioka and Suzuki, 2009) that imposes a mixture of ℓ_1-mixed-norm and ℓ_2-mixed-norm regularizations. Numerical studies have shown that ℓ_p-MKL with $p > 1$ and elasticnet-MKL show better performances than ℓ_1-MKL in several situations (Kloft et al., 2009; Cortes et al., 2009b; Tomioka and Suzuki, 2009). An interesting notion here is that both ℓ_p-MKL and elasticnet-MKL produce denser estimator than the original ℓ_1-MKL while they show favorable performances. The goal of this paper is to give a theoretical justification to these experimental results favorable for the *dense type* MKL methods. To this aim, we give a unifying framework to derive a fast learning rate of an *arbitrary* norm type regularization, and discuss which regularization is preferred depending on the problem settings.

In the pioneering paper of Lanckriet et al. (2004), a convergence rate of MKL was given as $\sqrt{M/n}$, where M is the number of given kernels and n is the sample size. Ying and Zhou (2007) introduced a uniform Glivenko-Cantelli class and analyzed learning a Gaussian kernel with flexible kernel width. This work was improved by Micchelli et al. (2016). Srebro and Ben-David (2006) gave simpler analysis for learning kernels based on the pseudo-dimension of the given kernel class. Ying and Campbell (2009) gave a convergence bound utilizing Rademacher chaos and gave some upper bounds of the Rademacher chaos utilizing the pseudo-dimension of the kernel class. Cortes et al. (2009b) presented a convergence bound for a learning method with L_2 regularization on the kernel weight. Cortes et al. (2010) gave the convergence rate of ℓ_p-MKL as $\sqrt{\log(M)/n}$ for $p = 1$ and $M^{1-\frac{1}{p}}/\sqrt{n}$ for $1 < p \leq 2$. Kloft et al. (2011) gave a similar convergence bound with improved constants. Kloft et al. (2010) generalized this bound to a variant of the elasticnet type regularization and widened the effective range of p to all range of $p \geq 1$ while $1 \leq p \leq 2$ had been imposed in the existing works. One concern about these bounds is that all bounds introduced above are "global" bounds in a sense that the bounds are applicable to all candidates of estimators. Consequently all convergence rate presented above are of order $1/\sqrt{n}$ with respect to the number n of samples. However, by utilizing the *localization* techniques including so-called local Rademacher complexity

(Bartlett et al., 2005; Koltchinskii, 2006) and peeling device (van de Geer, 2000), we can derive a faster learning rate. Instead of uniformly bounding all candidates of estimators, the localized inequality focuses on a particular estimator such as empirical risk minimizer, thus can give a sharp convergence rate. As for a kernel learning, Wu et al. (2007) derived a fast learning rate for classification problem with kernel parameter optimization, but it does not give explicit analysis for learning linear combination of kernels with convex regularization as in ℓ_p-MKL.

Localized bounds of MKL have been given mainly in sparse learning settings (Koltchinskii and Yuan, 2008; Meier et al., 2009; Koltchinskii and Yuan, 2010), and there are only few studies for non-sparse settings in which the sparsity of the ground truth is not assumed. The first localized bound of MKL is derived by Koltchinskii and Yuan (2008) in the setting of ℓ_1-MKL. The second one was given by Meier et al. (2009) who gave a near optimal convergence rate for elasticnet type regularization. Koltchinskii and Yuan (2010) considered a variant of ℓ_1-MKL and showed it achieves the minimax optimal convergence rate. All these localized convergence rates were considered in sparse learning settings, and it has not been discussed how a dense type regularization outperforms the sparse ℓ_1-regularization. Kloft and Blanchard (2011) gave a localized convergence bound of ℓ_p-MKL. However, their analysis assumed a strong condition where RKHSs have no-correlation to each other. Suzuki and Sugiyama (2013) gave a fast learning rate of the elasticnet regularization and discussed the difference between ℓ_1-regularization and the elasticnet regularization based on the smoothness of the true function.

In this paper, we show a unifying framework to derive fast convergence rates of MKL with various regularization types. The framework is applicable to *arbitrary* mixed-norm regularizations including ℓ_p-MKL and elasticnet-MKL. Our learning rate utilizes the localization technique, thus is tighter than global type learning rates. We discuss our bound in two situations: *homogeneous complexity* situation and *inhomogeneous complexity* situation where homogeneous complexity means that all RKHSs have the same *complexities* and inhomogeneous complexity means that the complexities of RKHSs are different to each other. In the homogeneous situation, we apply our general framework to some examples and show our bound achieves the minimax-optimal rate. As a by-product, we obtain a tighter convergence rate of ℓ_p-MKL than existing results. Moreover we show that our bound indicates that ℓ_1-MKL shows the best performance among all "isotropic" mixed-norm regularizations in homogeneous settings. Next we analyze our bound in inhomogeneous settings where the *complexities* of the RKHSs are not uniformly same. We show that dense type regularizations can give better generalization error bounds than the sparse ℓ_1-regularization in the inhomogeneous setting. Here it should be noted that in real settings inhomogeneous complexity is more natural than homogeneous complexity. Finally we give numerical experiments to show the validity of the theoretical investigations. We see that the numerical experiments well support the theoretical findings. As far as the author knows, this is the first theoretical attempt to clearly show the inhomogeneous complexities are advantageous for dense type MKL.

2. Preliminary

In this section we give the problem formulation, the notations and the assumptions required for the convergence analysis.

2.1. Problem formulation

Suppose that we are given n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ distributed from a probability distribution P on $\mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is an input space. We denote by Π the marginal distribution of P on $\mathcal{X}$. We are given M reproducing kernel Hilbert spaces (RKHS) $\{\mathcal{H}_m\}_{m=1}^M$ each of which is associated with a kernel k_m. We consider a mixed-norm type regularization with respect to an arbitrary given norm $\|\cdot\|_\psi$, that is, the regularization is given by the norm $\|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$ of the vector $(\|f_m\|_{\mathcal{H}_m})_{m=1}^M$ for $f_m \in \mathcal{H}_m$ $(m = 1, \ldots, M)$[1]. For notational simplicity, we write $\|f\|_\psi = \|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$ for $f = \sum_{m=1}^M f_m$ $(f_m \in \mathcal{H}_m)$.

The general formulation of MKL, we consider in this paper, fits a function $f = \sum_{m=1}^M f_m$ $(f_m \in \mathcal{H}_m)$ to the data by solving the following optimization problem:

$$\hat{f} = \sum_{m=1}^M \hat{f}_m = \operatorname*{arg\,min}_{f_m \in \mathcal{H}_m \ (m=1,\ldots,M)} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{m=1}^M f_m(x_i) \right)^2 + \lambda_1^{(n)} \|f\|_\psi^2. \quad (1)$$

We call this "ψ-norm MKL". This formulation covers many practically used MKL methods (e.g., ℓ_p-MKL, elasticnet-MKL, variable sparsity kernel learning (see later for their definitions)), and is solvable by a finite dimensional optimization procedure due to the representer theorem (Kimeldorf and Wahba, 1971). In this paper, we mainly focus on the regression problem (the squared loss). However the discussion can be generalized to Lipschitz continuous and strongly convex losses as in Bartlett et al. (2005) (see Section 7).

Example 1: ℓ_p-MKL The first motivating example of ψ-norm MKL is ℓ_p-MKL (Kloft et al., 2009) that employs ℓ_p-norm for $1 \le p \le \infty$ as the regularizer: $\|f\|_\psi = \|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_{\ell_p} = (\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p)^{\frac{1}{p}}$. If p is strictly greater than 1 $(p > 1)$, the solution of ℓ_p-MKL becomes dense. In particular, $p = 2$ corresponds to averaging candidate kernels with uniform weight (Micchelli and Pontil, 2005). It is reported that ℓ_p-MKL with p greater than 1, say $p = \frac{4}{3}$, often shows better performance than the original sparse ℓ_1-MKL (Cortes et al., 2010).

Example 2: elasticnet-MKL The second example is elasticnet-MKL (Shawe-Taylor, 2008; Tomioka and Suzuki, 2009) that employs mixture of ℓ_1 and ℓ_2 norms as the regularizer: $\|f\|_\psi = \tau\|f\|_{\ell_1} + (1-\tau)\|f\|_{\ell_2} = \tau \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m} +$

[1] We assume that the mixed-norm $\|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$ satisfies the triangular inequality with respect to $(f_m)_{m=1}^M$, that is, $\|(\|f_m + f'_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi \le \|(\|f_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi + \|(\|f'_m\|_{\mathcal{H}_m})_{m=1}^M\|_\psi$. To satisfy this condition, it is sufficient if the norm is monotone, i.e., $\|\boldsymbol{a}\|_\psi \le \|\boldsymbol{a} + \boldsymbol{b}\|_\psi$ for all $\boldsymbol{a}, \boldsymbol{b} \ge \boldsymbol{0}$.

$(1-\tau)(\sum_{m=1}^{M}\|f_m\|_{\mathcal{H}_m}^2)^{\frac{1}{2}}$ with $\tau \in [0,1]$. Elasticnet-MKL shares the same spirit with ℓ_p-MKL in a sense that it bridges sparse ℓ_1-regularization and dense ℓ_2-regularization. Efficient optimization method for elasticnet-MKL is proposed by Suzuki and Tomioka (2011).

Example 3: Variable Sparsity Kernel Learning Variable Sparsity Kernel Learning (VSKL) proposed by Aflalo et al. (2011) divides the RKHSs into M' groups $\{\mathcal{H}_{j,k}\}_{k=1}^{M_j}$, $(j = 1,\dots,M')$ and imposes a mixed norm regularization $\|f\|_\psi = \|f\|_{(p,q)} = \left\{\sum_{j=1}^{M'}(\sum_{k=1}^{M_j}\|f_{j,k}\|_{\mathcal{H}_{j,k}}^p)^{\frac{q}{p}}\right\}^{\frac{1}{q}}$ where $1 \le p$, $1 \le q$, and $f_{j,k} \in \mathcal{H}_{j,k}$. An advantageous point of VSKL is that by adjusting the parameters p and q, various levels of sparsity can be introduced. The parameters can control the level of sparsity *within* group and *between* groups. This point is beneficial especially for multi-modal tasks like object categorization.

2.2. Notations and assumptions

Here, we prepare notations and assumptions that are used in the analysis. Let $\mathcal{H}^{\oplus M} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_M$. We utilize the same notation $f \in \mathcal{H}^{\oplus M}$ indicating both the vector $(f_1,\dots,f_M)$ and the function $f = \sum_{m=1}^{M} f_m$ $(f_m \in \mathcal{H}_m)$. This is a little abuse of notation because the decomposition $f = \sum_{m=1}^{M} f_m$ might not be unique as an element of $L_2(\Pi)$ in general. However, we will assume an *incoherence assumption* (Assumption 4), and under this assumption, the decomposition is unique. Therefore, this notation will not cause any confusion in this article.

Throughout the paper, we assume the following technical conditions (see also Bach (2008)).

Assumption 1. (Realizable Assumption)

(A1) *There exists* $f^* = (f_1^*,\dots,f_M^*) \in \mathcal{H}^{\oplus M}$ *such that* $\mathrm{E}[Y|X] = f^*(X) = \sum_{m=1}^{M} f_m^*(X)$, *and the noise* $\epsilon := Y - f^*(X)$ *is bounded as* $|\epsilon| \le L$ *or is a normal variable with mean* 0 *and variance* L^2 *(*$\epsilon \sim N(0, L^2)$*).*

Assumption 2. (Kernel Assumption)

(A2) *For each* $m = 1,\dots,M$, $\mathcal{H}_m$ *is separable (with respect to the RKHS norm) and* $\sup_{X\in\mathcal{X}}|k_m(X,X)| \le 1$.

The first assumption in (A1) ensures the model $\mathcal{H}^{\oplus M}$ is correctly specified, and the technical assumption $|\epsilon| \le L$ allows ϵf to be Lipschitz continuous with respect to f.

Let an integral operator $T_{k_m} : L_2(\Pi) \to L_2(\Pi)$ corresponding to a kernel function k_m be

$$T_{k_m} f = \int k_m(\cdot, x) f(x) \mathrm{d}\Pi(x).$$

It is known that this operator is compact, positive, and self-adjoint (see Theorem 4.27 of Steinwart (2008)). Thus it has at most countably many non-negative

eigenvalues. We denote by $\mu_{\ell,m}$ be the ℓ-th largest eigenvalue (with possible multiplicity) of the integral operator T_{k_m}. By Theorem 4.27 of Steinwart (2008), the sum of $\mu_{\ell,m}$ is bounded ($\sum_\ell \mu_{\ell,m} < \infty$), and thus $\mu_{\ell,m}$ decreases with order ℓ^{-1} ($\mu_{\ell,m} = o(\ell^{-1})$). We further assume the sequence of the eigenvalues converges even faster to zero.

Assumption 3. (Spectral Assumption) *There exist* $0 < s_m < 1$ *and* $0 < c$ *such that*

$$\mu_{\ell,m} \leq c\ell^{-\frac{1}{s_m}}, \quad (\forall \ell \geq 1,\ 1 \leq \forall m \leq M), \tag{A3}$$

where $\{\mu_{\ell,m}\}_{\ell=1}^{\infty}$ *is the spectrum of the operator* T_{k_m} *corresponding to the kernel* k_m*. Otherwise, the RKHS* $\mathcal{H}_m$ *is finite dimensional with a bounded dimension;* $s_m = 0$ *and there exists* $c > 0$ *such that*

$$\mu_{\ell,m} > 0 \quad (\forall \ell \leq c), \quad \mu_{\ell,m} = 0 \quad (\forall \ell > c).$$

It was shown that the spectral assumption (A3) is equivalent to the classical covering number assumption (Steinwart et al., 2009). Recall that the ϵ-covering number $N(\epsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi))$ with respect to $L_2(\Pi)$ is the minimal number of balls with radius ϵ needed to cover the unit ball $\mathcal{B}_{\mathcal{H}_m}$ in $\mathcal{H}_m$ (van der Vaart and Wellner, 1996). If the spectral assumption (A3) and the boundedness assumption (A2) holds, there exists a constant C that depends only on s and c such that

$$\log N(\varepsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi)) \leq C\varepsilon^{-2s_m}, \tag{2}$$

and the converse is also true (see Steinwart et al. (2009, Theorem 15) and Steinwart (2008) for details). Therefore, if s_m is large, the RKHSs are regarded as "complex", and if s_m is small, the RKHSs are "simple".

An important class of RKHSs where s_m is known is Sobolev space. (A3) holds with $s_m = \frac{d}{2\alpha}$ for Sobolev space $W^{\alpha,2}(\mathcal{X})$ of α-times continuously differentiability on the Euclidean ball $\mathcal{X}$ of $\mathbb{R}^d$ (Edmunds and Triebel, 1996). Moreover, for α-times differentiable kernels on a closed Euclidean ball in $\mathbb{R}^d$, (A3) holds for $s_m = \frac{d}{2\alpha}$ (Steinwart, 2008, Theorem 6.26). According to Theorem 7.34 of Steinwart (2008), for Gaussian kernels with compact support distribution, that holds for arbitrary small $0 < s_m$. The covering number of Gaussian kernels with *unbounded* support distribution is also described in Theorem 7.34 of Steinwart (2008).

When $s_m = 0$, the RKHS $\mathcal{H}_m$ is finite dimensional because of Assumption 3. The uniform boundedness of the dimensions of the RKHSs could be relaxed trivially, but we do not go into details of that direction for theoretical simplicity.

Let κ_M be defined as follows:

$$\kappa_M := \sup\left\{\kappa \geq 0 \;\middle|\; \kappa \leq \frac{\|\sum_{m=1}^M f_m\|_{L_2(\Pi)}^2}{\sum_{m=1}^M \|f_m\|_{L_2(\Pi)}^2},\ \forall f_m \in \mathcal{H}_m\ (m = 1, \ldots, M)\right\}. \tag{3}$$

κ_M represents the correlation of RKHSs. We assume all RKHSs are not completely correlated to each other.

TABLE 1
Summary of the constants we use in this article.

n	The number of samples.
M	The number of candidate kernels.
L	The bound of the noise (A2).
c	The coefficient for Spectral Assumption; see (A3).
s_m	The decay rate of spectrum; see (A3).
κ_M	The smallest eigenvalue of the design matrix; see Eq. (3).
C_1	The coefficient for Embedded Assumption; see (A5).

Assumption 4. (Incoherence Assumption) *κ_M is strictly bounded from below; there exists a constant $C_0 > 0$ such that*

$$0 < C_0^{-1} < \kappa_M. \tag{A4}$$

This condition is motivated by the *incoherence condition* (Koltchinskii and Yuan, 2008; Meier et al., 2009) considered in sparse MKL settings. This ensures the uniqueness of the decomposition $f^* = \sum_{m=1}^M f_m^*$ of the ground truth. This can be easily checked as follows: for different decompositions $f = \sum_{m=1}^M f_m = \sum_{m=1}^M g_m$, it holds that $0 = \|f - f\|_{L_2(\Pi)}^2 = \|\sum_{m=1}^M (f_m - g_m)\|_{L_2(\Pi)}^2 \geq \kappa \sum_{m=1}^M \|f_m - g_m\|_{L_2(\Pi)}^2$, then $f_m = g_m$ for all m. Bach (2008) also assumed this condition to show the consistency of ℓ_1-MKL.

Finally we give a technical assumption with respect to ∞-norm.

Assumption 5. (Embedded Assumption) *Under the Spectral Assumption, there exists a constant $C_1 > 0$ such that*

$$\|f_m\|_\infty \leq C_1 \|f_m\|_{\mathcal{H}_m}^{1-s_m} \|f_m\|_{L_2(\Pi)}^{s_m}. \tag{A5}$$

This condition is met when the input distribution Π has a density with respect to the uniform distribution on $\mathcal{X}$ that is bounded away from 0 and the RKHSs are continuously embedded in a Sobolev space $W^{\alpha,2}(\mathcal{X})$ where $s_m = \frac{d}{2\alpha}$, d is the dimension of the input space $\mathcal{X}$ and α is the "smoothness" of the Sobolev space. Many practically used kernels satisfy this condition (A5). For example, the RKHSs of Gaussian kernels can be embedded in all Sobolev spaces. Therefore the condition (A5) seems rather common and practical. More generally, there is a clear characterization of the condition (A5) in terms of *real interpolation of spaces*. One can find detailed and formal discussions of interpolations in Steinwart et al. (2009), and Proposition 2.10 of Bennett and Sharpley (1988) gives the necessary and sufficient condition for the assumption (A5).

Constants we use later are summarized in Table 1.

3. Convergence rate of ψ-norm MKL

Here we derive the learning rate of ψ-norm MKL in the most general setting. We suppose that the number of kernels M can increase along with the number of samples n. The motivation of our analysis is summarized as follows:

- Give a unifying framework to derive a sharp convergence rate of ψ-norm MKL.
- (homogeneous complexity) Show the convergence rate of some examples using our general framework, prove its minimax-optimality, and show the optimality of ℓ_1-regularization under conditions that the complexities s_m of all RKHSs are same.
- (inhomogeneous complexity) Discuss how the dense type regularization outperforms sparse type regularization, when the complexities s_m of all RKHSs are *not* uniformly same.

We define

$$\eta(t) := \eta_n(t) = \max(1, \sqrt{t}, t/\sqrt{n}),$$

for $t > 0$. For given positive reals $\{r_m\}_{m=1}^M$ and given n, we define $\alpha_1, \alpha_2, \beta_1, \beta_2$ as follows:

$$\alpha_1 := \alpha_1(\{r_m\}) = 3\left(\sum_{m=1}^M \frac{r_m^{-2s_m}}{n}\right)^{\frac{1}{2}}, \quad \alpha_2 := \alpha_2(\{r_m\}) = 3\left\|\left(\frac{s_m r_m^{1-s_m}}{\sqrt{n}}\right)_{m=1}^M\right\|_{\psi^*},$$

$$\beta_1 := \beta_1(\{r_m\}) = 3\left(\sum_{m=1}^M \frac{r_m^{-\frac{2s_m(3-s_m)}{1+s_m}}}{n^{\frac{2}{1+s_m}}}\right)^{\frac{1}{2}},$$

$$\beta_2 := \beta_2(\{r_m\}) = 3\left\|\left(\frac{s_m r_m^{\frac{(1-s_m)^2}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\right)_{m=1}^M\right\|_{\psi^*}, \tag{4}$$

(note that $\alpha_1, \alpha_2, \beta_1, \beta_2$ implicitly depends on the reals $\{r_m\}_{m=1}^M$). Then the following theorem gives the general form of the learning rate of ψ-norm MKL.

Theorem 1. *Suppose Assumptions 1-5 are satisfied. Let $\{r_m\}_{m=1}^M$ be arbitrary positive reals that can depend on n, and assume $\lambda_1^{(n)} \geq \left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2$. Then there exists a constant ϕ depending only on $\{s_m\}_{m=1}^M$, c, C_1, L such that for all n and t' that satisfy $\frac{\log(M)}{\sqrt{n}} \leq 1$ and $\frac{4\phi\sqrt{n}}{\kappa_M}\max\{\alpha_1^2, \beta_1^2, \frac{M\log(M)}{n}\}\eta(t') \leq \frac{1}{12}$ and for all $t \geq 1$, we have*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{24\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2 + \beta_1^2 + \frac{M\log(M)}{n}\right) + 4\lambda_1^{(n)}\|f^*\|_\psi^2, \tag{5}$$

with probability $1-\exp(-t)-\exp(-t')$. In particular, for $\lambda_1^{(n)} = \left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2$, we have

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{24\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2 + \beta_1^2 + \frac{M\log(M)}{n}\right) + 4\left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2. \tag{6}$$

The proof will be given in Appendix C. The statement of Theorem 1 itself is complicated. Thus we will show later concrete learning rates on some examples such as ℓ_p-MKL. The convergence rate (6) depends on the positive reals $\{r_m\}_{m=1}^M$, but the choice of $\{r_m\}_{m=1}^M$ are arbitrary. Thus by minimizing the right hand side of Eq. (6), we obtain tight convergence bound as follows:

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(\min_{\substack{\{r_m\}_{m=1}^M: \\ r_m>0}} \left\{\alpha_1^2 + \beta_1^2 + \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2 + \frac{M\log(M)}{n}\right\}\right). \tag{7}$$

There is a trade-off between the first two terms $(a) := \alpha_1^2 + \beta_1^2$ and the third term $(b) := \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2$, that is, if we take $\{r_m\}_m$ large, then the term (a) becomes small and the term (b) becomes large, on the other hand, if we take $\{r_m\}_m$ small, then it results in large (a) and small (b). Therefore we need to balance the two terms (a) and (b) to obtain the minimum in Eq. (7).

We discuss the obtained learning rate in two situations, (i) *homogeneous complexity* situation, and (ii) *inhomogeneous complexity* situation:
(i) (homogeneous) All s_ms are same: there exists $0 < s < 1$ such that $s_m = s$ $(\forall m)$ (Sec.4).
(ii) (inhomogeneous) All s_ms are *not* same: there exist m, m' such that $s_m \neq s_{m'}$ (Sec.5).

4. Analysis on homogeneous settings

Here we assume all s_ms are same, say $s_m = s$ for all m (homogeneous setting). In this section, we give a simple upper bound of the minimum of the bound (7) (Sec.4.1), derive concrete convergence rates of some examples using the simple upper bound (Sec.4.2) and show that the simple upper bound achieves the minimax learning rate of ψ-norm ball if ψ-norm is isotropic (Sec.4.3). Finally we discuss the optimal regularization (Sec.4.4). In Sec.4.2, we also discuss the difference between our bound of ℓ_p-MKL and existing bounds.

4.1. Simplification of convergence rate

If we restrict the situation as all r_ms are same ($r_m = r$ $(\forall m)$ for some r), then the minimization in Eq. (7) can be easily carried out as in the following lemma. Let $\mathbf{1}$ be the M-dimensional vector each element of which is 1: $\mathbf{1} := (1, \ldots, 1)^\top \in \mathbb{R}^M$, and $\|\cdot\|_{\psi^*}$ be the dual norm of the ψ-norm[2].

Lemma 2. *Suppose* $s_m = s$ $(\forall m)$ *with some* $0 < s < 1$, *and set* $\lambda_1^{(n)} = 18M^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} \|\mathbf{1}\|_{\psi^*}^{\frac{2s}{1+s}} \|f^*\|_\psi^{-\frac{2}{1+s}}$, *then for all* n *and* t' *that sat-*

[2]The dual of the norm $\|\cdot\|_\psi$ is defined as $\|\boldsymbol{b}\|_{\psi^*} := \sup_{\boldsymbol{a}}\{\boldsymbol{b}^\top\boldsymbol{a} \mid \|\boldsymbol{a}\|_\psi \leq 1\}$.

isfy $\frac{4\phi}{\kappa_M}\left\{9\left(\frac{M}{\sqrt{n}}\right)^{\frac{1-s}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} \vee \frac{M\log(M)}{\sqrt{n}}\right\}\eta(t') \leq \frac{1}{12}$ *and* $n \geq (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi/M)^{\frac{4s}{1-s}}$*, and for all* $t \geq 1$*, we have*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq C\eta(t)^2\left\{M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right\},$$

with probability $1-\exp(-t)-\exp(-t')$ *where* C *is a constant depending on* ϕ *and* κ_M*. In particular we have*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left\{M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right\}. \tag{8}$$

The proof is given in Appendix F.1. Lemma 2 is derived by assuming $r_m = r$ $(\forall m)$, which might make the bound loose. However, when the norm $\|\cdot\|_\psi$ is *isotropic* (whose definition will appear later), that restriction ($r_m = r$ $(\forall m)$) does not make the bound loose, that is, the upper bound obtained in Lemma 2 is tight and achieves the minimax optimal rate (the minimax optimal rate is the one that cannot be improved by any estimator). In the following, we investigate the general result of Lemma 2 through some important examples.

4.2. Convergence rate of some examples

4.2.1. Convergence rate of ℓ_p-MKL

Here we derive the convergence rate of ℓ_p-MKL ($1 \leq p \leq \infty$) where $\|f\|_\psi = \sum_{m=1}^M(\|f_m\|_{\mathcal{H}_m}^p)^{\frac{1}{p}}$ (for $p=\infty$, it is defined as $\max_m \|f_m\|_{\mathcal{H}_m}$). It is well known that the dual norm of ℓ_p-norm is given as ℓ_q-norm where q is the real satisfying $\frac{1}{p}+\frac{1}{q}=1$. For notational simplicity, let $R_p := \left(\sum_{m=1}^M \|f_m^*\|_{\mathcal{H}_m}^p\right)^{\frac{1}{p}}$. Then substituting $\|f^*\|_\psi = R_p$ and $\|\mathbf{1}\|_{\psi^*} = \|\mathbf{1}\|_{\ell_q} = M^{\frac{1}{q}} = M^{1-\frac{1}{p}}$ into the bound (8), the learning rate of ℓ_p-MKL is given as

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}}M^{1-\frac{2s}{p(1+s)}}R_p^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right). \tag{9}$$

If we further assume n is sufficiently large such that

$$n \geq M^{\frac{2}{p}}R_p^{-2}(\log M)^{\frac{1+s}{s}}, \tag{10}$$

then the leading term is the first term, and thus we have

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}}M^{1-\frac{2s}{p(1+s)}}R_p^{\frac{2s}{1+s}}\right). \tag{11}$$

Note that as the complexity s of RKHSs becomes small the convergence rate becomes fast. It is known that $n^{-\frac{1}{1+s}}$ is the minimax optimal learning rate for

single kernel learning. The derived rate of ℓ_p-MKL is obtained by multiplying a coefficient depending on M and R_p to the optimal rate of single kernel learning. To investigate the dependency of R_p to the learning rate, let us consider two extreme settings, i.e., sparse setting $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = (1, 0, \ldots, 0)$ and dense setting $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = (1, \ldots, 1)$ as in Kloft et al. (2011).

- $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = (1, 0, \ldots, 0)$: $R_p = 1$ for all p. Therefore the convergence rate $n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}}$ is fast for small p and the minimum is achieved at $p = 1$. This means that ℓ_1 regularization is preferred for sparse truth.
- $(\|f_m^*\|_{\mathcal{H}_m})_{m=1}^M = (1, \ldots, 1)$: $R_p = M^{\frac{1}{p}}$, thus the convergence rate is $Mn^{-\frac{1}{1+s}}$ for all p. Interestingly for dense ground truth, there is no dependency of the convergence rate on the parameter p (later we will show that this is not the case in inhomogeneous setting (Sec.5)). That is, the convergence rate is M times the optimal learning rate of single kernel learning $(n^{-\frac{1}{1+s}})$ for all p. This means that for the dense settings, the complexity of solving MKL problem is equivalent to that of solving M single kernel learning problems.

Comparison with existing bounds Here we compare the bound for ℓ_p-MKL we derived above with the existing bounds. Let $\mathcal{H}_{\ell_p}(R_p)$ be the ℓ_p-mixed norm ball with radius R_p: $\mathcal{H}_{\ell_p}(R_p) := \{f = \sum_{m=1}^M f_m \mid (\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p)^{\frac{1}{p}} \leq R_p\}$. There are two types of convergence rates: global bound and localized bound.

(comparison with existing global bound) Cortes et al. (2010); Kloft et al. (2010, 2011) gave "global" type bounds for ℓ_p-MKL as

$$R(f) \leq \widehat{R}(f) + C \begin{cases} \sqrt{\frac{\log(M)}{n}} R_p & (p = 1), \\ \frac{M^{1-\frac{1}{p}}}{\sqrt{n}} R_p & (p > 1), \end{cases} \quad (\text{for all } f \in \mathcal{H}_{\ell_p}(R_p)), \tag{12}$$

where $R(f)$ and $\widehat{R}(f)$ is the population risk and the empirical risk. The bounds by Cortes et al. (2010) and Kloft et al. (2011) are restricted to the situation $1 \leq p \leq 2$. On the other hand, our analysis and that of Kloft et al. (2010) covers all $p \geq 1$.

Since our bound is specialized to the regularized risk minimizer $\hat{f}$ defined at Eq. (1) while the existing bound (12) is applicable to all $f \in \mathcal{H}_{\ell_p}(R_p)$, our bound is sharper than theirs for sufficiently large n. To see this, suppose that

$$n \geq \begin{cases} M^2 R_1^{-2} (\log M)^{-\frac{1+s}{1-s}} & (p = 1), \\ M^{\frac{2}{p}} R_p^{-2} & (p > 1), \end{cases} \tag{13}$$

then we have $n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p(1+s)}} R_p^{\frac{2s}{1+s}} \leq n^{-\frac{1}{2}} (M^{1-\frac{1}{p}} \vee \log(M)) R_p$ and hence our localized bound is sharper than the global one. Interestingly, the range of n presented in Eq. (13) where the localized bound exceeds the global bound is same

(up to $\log M$ term) as the range presented in Eq. (10) ($n \geq M^{\frac{2}{p}} R_p^{-2} (\log M)^{\frac{1+s}{s}}$) where the first term in our bound (9) dominates its second term so that the simplified bound (11) holds. That means that, at the "phase transition point" from global to localized bound, the first informative term in our bound becomes the leading term.

Finally we note that, since s can be large as long as Spectral Assumption (A3) is satisfied, the bound (12) is recovered by our analysis by approaching s to 1.

(comparison with existing localized bound) Kloft and Blanchard (2011) gave a tighter convergence rate utilizing the localization technique as

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\Big(\min_{p' \geq p} \Big\{ \tfrac{p'}{p'-1} n^{-\frac{1}{1+s}} M^{1-\frac{2s}{p'(1+s)}} R_{p'}^{\frac{2s}{1+s}} \Big\} \Big), \tag{14}$$

under a strong condition $\kappa_M = 1$ that imposes all RKHSs are completely uncorrelated to each other. Comparing our bound with their result, there is $\min_{p' \geq p}$ and $\frac{p'}{p'-1}$ in their bound (if there is not the term $\frac{p'}{p'-1}$, then the minimum of $\min_{p' \geq p}$ is attained at $p' = p$, thus our bound is tighter). Due to this, we obtain a quite different consequence from theirs. According to our bound (11), the optimal regularization among all ℓ_p-norm that gives the smallest generalization error is ℓ_1-regularization (this will be discussed later in Sec.4.4) while their consequence says that the optimal p changes depending on the "sparsity" of the true function f^*. Moreover we will observe that ℓ_1-regularization is optimal among *all* isotropic mixed-norm-type regularization. The details of the optimality will be discussed in Sec.4.4.

4.2.2. Convergence rate of elasticnet-MKL

Elasticnet-MKL employs a mixture of ℓ_1 and ℓ_2 norm as the regularizer:

$$\|f\|_\psi = \tau \|f\|_{\ell_1} + (1-\tau)\|f\|_{\ell_2}$$

where $\tau \in [0,1]$.

Then its dual norm is given by $\|\boldsymbol{b}\|_{\psi *} = \min_{\boldsymbol{a} \in \mathbb{R}^M} \left\{ \max\left(\frac{\|\boldsymbol{a}\|_{\ell_\infty}}{\tau}, \frac{\|\boldsymbol{a}-\boldsymbol{b}\|_{\ell_2}}{1-\tau} \right) \right\}$. Therefore by a simple calculation, we have $\|\mathbf{1}\|_{\psi *} = \frac{\sqrt{M}}{1-\tau+\tau\sqrt{M}}$. Hence Eq. (8) gives the convergence rate of elasticnet-MKL as

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} \frac{M^{1-\frac{s}{1+s}}}{(1-\tau+\tau\sqrt{M})^{\frac{2s}{1+s}}} (\tau\|f^*\|_{\ell_1} + (1-\tau)\|f^*\|_{\ell_2})^{\frac{2s}{1+s}} + \frac{M\log(M)}{n} \right).$$

Note that, when $\tau = 0$ or $\tau = 1$, this rate is identical to that of ℓ_2-MKL or ℓ_1-MKL obtained in Eq. (9) respectively.

4.2.3. Convergence rate of VSKL

Variable Sparsity Kernel Learning (VSKL) employs a mixed norm regularization defined by

$$\|f\|_\psi = \|f\|_{(p,q)} = \left\{ \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} \|f_{j,k}\|_{\mathcal{H}_{j,k}}^p \right)^{\frac{q}{p}} \right\}^{\frac{1}{q}},$$

where RKHSs are divided into M' groups $\{\mathcal{H}_{j,k}\}_{k=1}^{M_j}$, $(j = 1, \ldots, M')$ and $1 \leq p,\ 1 \leq q$.

Lemma 3. *The dual of the mixed norm is given by*

$$\|\boldsymbol{b}\|_{\psi^*} = \left\{ \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} |b_{j,k}|^{p^*} \right)^{\frac{q^*}{p^*}} \right\}^{\frac{1}{q^*}},$$

for $b_{j,k} \in \mathbb{R}$ $(k = 1, \ldots, M_j,\ j = 1, \ldots, M')$.

The proof will be given in Appendix F.2. Therefore the dual norm of the vector $\mathbf{1}$ is given by $\|\mathbf{1}\|_{\psi^*} = \left(\sum_{j=1}^{M'} M_j^{\frac{q^*}{p^*}} \right)^{\frac{1}{q^*}}$. Hence, by Eq. (8), the convergence rate of VSKL is given as

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\
=&\mathcal{O}_p\left(n^{-\frac{1}{1+s}} \left(\sum_{j=1}^{M'} M_j \right)^{1-\frac{2s}{1+s}} \left[\left(\sum_{j=1}^{M'} M_j^{\frac{q^*}{p^*}} \right)^{\frac{1}{q^*}} \left\{ \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} \|f_{j,k}^*\|_{\mathcal{H}_{j,k}}^p \right)^{\frac{q}{p}} \right\}^{\frac{1}{q}} \right]^{\frac{2s}{1+s}} \right. \\
&\left. + \frac{M \log(M)}{n} \right).
\end{aligned}$$

One can check that this convergence rate coincides with that of ℓ_p-MKL when $M' = 1$.

4.3. Minimax lower bound

In this section, we show that the derived learning rate (8) achieves the minimax-learning rate on the ψ-norm ball

$$\mathcal{H}_\psi(R) := \left\{ f = \sum_{m=1}^{M} f_m \;\middle|\; \|f\|_\psi \leq R \right\},$$

when the norm is *isotropic*.

Definition 4. *We say that ψ-norm $\|\cdot\|_\psi$ is isotropic when there exits a universal constant $\bar{c}$ such that*

$$\bar{c}M = \bar{c}\|\mathbf{1}\|_{\ell_1} \geq \|\mathbf{1}\|_{\psi^*}\|\mathbf{1}\|_\psi, \qquad \|\boldsymbol{b}\|_\psi \leq \|\boldsymbol{b}'\|_\psi \quad (\textit{if } 0 \leq b_m \leq b'_m\ (\forall m)), \tag{15}$$

(note that the inverse inequality $M \leq \|\mathbf{1}\|_{\psi^*}\|\mathbf{1}\|_\psi$ of the first condition always holds by the definition of the dual norm).

Practically used regularizations usually satisfy the isotropic property. In fact, ℓ_p-MKL, elasticnet-MKL and VSKL satisfy the isotropic property with $\bar{c} = 1$.

We derive the minimax learning rate in a simpler situation. First we assume that each RKHS is same as others. That is, the input vector is decomposed into M components like $x = (x^{(1)}, \ldots, x^{(M)})$ where $\{x^{(m)}\}_{m=1}^M$ are M i.i.d. copies of a random variable $\tilde{X}$, and $\mathcal{H}_m = \{f_m \mid f_m(x) = f_m(x^{(1)}, \ldots, x^{(M)}) = \tilde{f}_m(x^{(m)}),\ \tilde{f}_m \in \widetilde{\mathcal{H}}\}$ where $\widetilde{\mathcal{H}}$ is an RKHS shared by all $\mathcal{H}_m$. Thus $f \in \mathcal{H}^{\oplus M}$ is decomposed as $f(x) = f(x^{(1)}, \ldots, x^{(M)}) = \sum_{m=1}^M \tilde{f}_m(x^{(m)})$ where each $\tilde{f}_m$ is a member of the common RKHS $\widetilde{\mathcal{H}}$. We denote by $\tilde{k}$ the kernel associated with the RKHS $\widetilde{\mathcal{H}}$. We call this situation a *completely homogeneous model*.

In addition to the condition about the upper bound of spectrum (Spectral Assumption (A3)), we assume that the spectrum of all the RKHSs $\mathcal{H}_m$ have the same lower bound of polynomial rate.

Assumption 6. (Strong Spectral Assumption) *There exist* $0 < s < 1$ *and* $0 < c, c'$ *such that*

$$c' \ell^{-\frac{1}{s}} \leq \tilde{\mu}_\ell \leq c \ell^{-\frac{1}{s}}, \quad (1 \leq \forall \ell), \tag{A6}$$

where $\{\tilde{\mu}_\ell\}_{\ell=1}^\infty$ *is the spectrum of the integral operator* $T_{\tilde{k}}$ *corresponding to the kernel* $\tilde{k}$. *In particular, the spectrum of* T_{k_m} *also satisfies* $\mu_{\ell,m} \sim \ell^{-\frac{1}{s}}$ $(\forall \ell, m)$.

Without loss of generality, we may assume that $\mathrm{E}[f(\tilde{X})] = 0$ $(\forall f \in \widetilde{\mathcal{H}})$. Since each f_m receives i.i.d. copy of $\tilde{X}$, $\mathcal{H}_m$s are orthogonal to each other:

$$\begin{aligned} &\mathrm{E}[f_m(X) f_{m'}(X)] = \mathrm{E}[\tilde{f}_m(X^{(m)}) \tilde{f}_{m'}(X^{(m')})] = 0 \\ &\quad (\forall f_m \in \mathcal{H}_m,\ \forall f_{m'} \in \mathcal{H}_{m'},\ 1 \leq \forall m \neq m' \leq M). \end{aligned} \tag{16}$$

We also assume that the noise $\{\epsilon_i\}_{i=1}^n$ is an i.i.d. standard normal sequence.

Under the assumptions described above, we have the following minimax $L_2(\Pi)$-error.

Theorem 5. *Suppose* $R > 0$ *is given and* $n > \frac{\bar{c}^2 M^2}{R^2 \|\mathbf{1}\|_{\psi^*}^2}$ *is satisfied. We suppose the completely homogeneous model, and assume that Assumption 6 holds, the condition (16) is satisfied, and the noise* $\{\epsilon_i\}_{i=1}^n$ *is an i.i.d. standard normal sequence. Then the minimax-learning rate on* $\mathcal{H}_\psi(R)$ *for isotropic norm* $\|\cdot\|_\psi$ *is lower bounded as*

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_\psi(R)} \mathrm{E}\left[\|\hat{f} - f^*\|_{L_2(\Pi)}^2\right] \geq C M^{1 - \frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*} R)^{\frac{2s}{1+s}}, \tag{17}$$

where inf *is taken over all measurable functions of* n *samples* $\{(x_i, y_i)\}_{i=1}^n$.

The proof will be given in Appendix E. One can see that the convergence rate derived in Eq. (8) achieves the minimax rate on the ψ-norm ball (Theorem 5) up to $\frac{M \log(M)}{n}$ that is negligible when the number of samples is large. Indeed if

$$n \geq \frac{M^2 \log(M)^{\frac{1+s}{s}}}{\|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_\psi^2}, \tag{18}$$

then the first term in Eq. (8) dominates the second term $\frac{M\log(M)}{n}$ and the upper bound coincides with the minimax optima rate. Note that the condition (18) for the sample size n is equivalent to the condition for n assumed in Theorem 5 up to factors of $\log(M)^{\frac{1+s}{s}}$ and a constant.

The fact that ψ-norm MKL achieves the minimax optimal rate (17) indicates that the ψ-norm regularization is well suited to make the estimator included in the ψ-norm ball.

4.4. Optimal regularization strategy

Here we discuss which regularization gives the best performance based on the generalization error bound given by Lemma 2. Surprisingly the best regularization that gives the optimal performance among *all isotropic ψ-norm* regularizations is ℓ_1-norm regularization. This can be seen as follows. According to Eq. (8), we have seen that the convergence rate of ψ-norm MKL is upper bounded as

$$\|\hat{f}-f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left\{M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_{\psi})^{\frac{2s}{1+s}} + \frac{M\log(M)}{n}\right\},$$

and this is mini-max optimal on ψ-norm ball if ψ-norm is isotropic. Here by the definition of the dual norm $\|\cdot\|_{\psi^*}$, we always have

$$\|f^*\|_{\ell_1} = \sum_{m=1}^M \|f_m^*\|_{\mathcal{H}_m} = \sum_{m=1}^M 1\times\|f_m^*\|_{\mathcal{H}_m} \le \|\mathbf{1}\|_{\psi^*}\|f^*\|_{\psi}. \tag{19}$$

Therefore the leading term of the convergence rate for ℓ_1-norm regularization is upper bounded by that for other arbitrary ψ-norm regularization as

$$M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}\|f^*\|_{\ell_1}^{\frac{2s}{1+s}} \le M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_{\psi})^{\frac{2s}{1+s}},$$

(here it should be noticed that the dual norm of ℓ_1-norm is ℓ_∞-norm and $\|\mathbf{1}\|_{\ell_\infty}=1$). This shows that the upper bound (8) is minimized by ℓ_1-norm regularization. In other words, ℓ_1-regularization is optimal among *all* (isotropic) ψ-norm regularization in homogeneous settings.

This consequence is different from that of Kloft and Blanchard (2011) where the optimal regularization among ℓ_p-MKL is discussed. Their consequence says that the best performance is achieved at $p\gneq 1$ and the best p depends on the variation of the RKHS norms of $\{f_m^*\}_{m=1}^M$: if f^* is close to sparse (i.e., $\|f_m^*\|_{\mathcal{H}_m}$ decays rapidly), small p is preferred, on the other hand if f^* is dense (i.e., $\{\|f_m^*\|_{\mathcal{H}_m}\}_{m=1}^M$ is uniform), then large p is preferred. This consequence seems reasonable, but our consequence is different: ℓ_1-norm regularization is always optimal in ℓ_p-regularizations. The antinomy of the two consequences comes from the additional terms $\min_{p'\ge p}$ and $\frac{p'}{p'-1}$ in their bound (14) (there are no such terms in our bound). This difference makes our bound tighter than their bound but simultaneously leads to a somewhat counter-intuitive consequence that is contrastive against the some experiment results supporting dense type regu-

larization. However such experimental observations are justified by considering *inhomogeneous settings*. Here we should notice that the homogeneous setting is quite restrictive and unrealistic because it is required that the complexities of all RKHSs are uniformly same. In real settings, it is natural to assume the complexities varies depending on RKHS (inhomogeneous). In the next section, we discuss how dense type regularizations outperform the ℓ_1-regularization.

5. Analysis on inhomogeneous settings

In the previous sections (analysis on homogeneous settings), we have seen ℓ_1-MKL shows the best performance among isotropic ψ-norm and have not observed any theoretical justification supporting the fact that dense MKL methods like $\ell_{\frac{4}{3}}$-MKL can outperform the sparse ℓ_1-MKL (Cortes et al., 2010). In this section, we show dense type regularizations can outperform the sparse regularization in inhomogeneous settings (where there exists m, m' such that $s_m \neq s_{m'}$). For simplicity, we focus on ℓ_p-MKL, and discuss the relation between the learning rate and the norm parameter p.

Let us consider an extreme situation where $s_1 = s_2 = \cdots = s_d = s$ for some $0 < s < 1$ and $1 \leq d \leq M$, and $s_m = 0$ $(m > d)$ under Assumption 3 (that is, all RKHSs $\mathcal{H}_m$ for $m \geq d$ are finite dimensional). In this situation, we may have

$$\alpha_1 = 3\left(\frac{dr_1^{-2s} + M - d}{n}\right)^{\frac{1}{2}}, \ \alpha_2 = 3\frac{sr_1^{1-s}}{\sqrt{n}}\|\mathbf{1}_d\|_{\psi^*},$$
$$\beta_1 = 3\left(\frac{dr_1^{-\frac{2s(3-s)}{1+s}} + M - d}{n^{\frac{2}{1+s}}}\right)^{\frac{1}{2}}, \ \beta_2 = 3\frac{sr_1^{\frac{(1-s)^2}{1+s}}}{n^{\frac{1}{1+s}}}\|\mathbf{1}_d\|_{\psi^*}.$$

for all norm ψ where $\mathbf{1}_d = (\underbrace{1, \ldots, 1}_{d \text{ elements}}, 0, \ldots, 0)^\top$. Note that these α_1, α_2, β_1 and β_2 have dependency on the choice of the norm $\|\cdot\|_\psi$. Through a bit cumbersome calculation, we obtain the following lemma that describes how the choice of the regularization affects the generalization error in an inhomogeneous setting. The lemma indicates that, in an inhomogeneous setting, ℓ_p-regularization with $1 < p < \infty$ could achieve better generalization than $p = 1$ and $p = \infty$. Here, we denote by $\hat{f}^{(p)}$ the trained function by ℓ_p-MKL.

Lemma 6. *Suppose that there exits $1 < s < 1$, and $1 \leq d \leq M$ such that $s_m = s$ for $1 \leq m \leq d$ and $s_m = 0$ $(m > d)$ under Assumption 3 and $\|f_m^*\|_{\mathcal{H}_m} \leq 1$ for all m. If $n \geq M^{\frac{4s}{1-s}} \vee (M \log(M))^{\frac{1+s}{s}}$ and $\|f^*\|_{\ell_p} \geq 1$ for all $p \geq 1$, then the bound (7) implies*

$$\|\hat{f}^{(p)} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} d^{1-\frac{2s}{p(1+s)}} \|f^*\|_{\ell_p}^{\frac{2s}{1+s}}\right).$$

In particular, if $d = M^{b_1}$ and $\|f_m^*\|_{\mathcal{H}_m} = m^{-b_2}$ for $0 < b_1 < 1$ and $0 < b_2 < 1$, then it holds that

$$\|\hat{f}^{(1)} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{\frac{2s}{1+s}(1-b_2)+\frac{1-s}{1+s}b_1}\right),$$

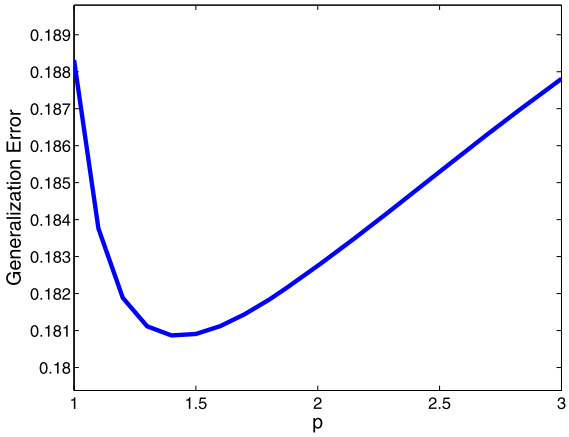


FIG 1. *The generalization error bound* (20) *of* ℓ_p*-MKL with respect to* p.

$$\|\hat{f}^{(\infty)} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{b_1}\right),$$
$$\|\hat{f}^{(b_2^{-1})} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{b_1(1-\frac{2sb_2}{1+s})} \log(M)^{b_2 \frac{2s}{1+s}}\right).$$

The proof is given in Appendix F.3. If d is very small so that $b_1 < \frac{2s}{1+s}(1-b_2)$, then it is easy to check that the bound for $\hat{f}^{(\infty)}$ is better than that of $\hat{f}^{(1)}$. This indicates that when the complexities of RKHSs are highly inhomogeneous, the generalization ability of *dense* type regularization (e.g., ℓ_∞-MKL) can be better than *sparse* type regularization (ℓ_1-MKL). Moreover, by the assumption $0 < b_1, b_2 < 1$, we can see that the bound for $\hat{f}^{(b_2^{-1})}$ is better than those of both $\hat{f}^{(1)}$ and $\hat{f}^{(\infty)}$. This indicates that the inhomogeneity of the complexities of the RKHSs and the norms of $(f_m^*)_m$ affects the optimal regularization. In particular, an *intermediate regularization* (ℓ_{1/b_2}-regularization) could be better than the extremal ones (ℓ_1 and ℓ_∞-regularizations).

Next we numerically calculate the convergence rate:

$$\min_{\substack{\{r_m\}_{m=1}^M: \\ r_m > 0}} \left\{ \alpha_1^2 + \beta_1^2 + \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2 \right] \|f^*\|_\psi^2 \right\}. \tag{20}$$

Here we randomly generated s_m from the uniform distribution on $[0, 1/3]$ and $\|f_m^*\|_{\mathcal{H}_m}$ from the uniform distribution on $[0, 1]$ with $n = 100$ and $M = 10$. Then calculated the minimum of Eq. (20) using a numerical optimization solver where ℓ_p-norm is employed as the regularizer (ℓ_p-MKL). We used *Differential Evolution* technique[3] (Lampinen, 2005; Chakraborty, 2008) to obtain the minimum value. Figure 1 plots the minimum value of Eq. (20) against the parameter p of ℓ_p-norm. We can see that the generalization error once goes down and then goes up as p gets large. The optimal p is attained around $p = 1.4$ in this example.

[3] We used the Matlab® code available in Chakraborty (2008).

In real settings, it is likely that one uses various types of kernels and the complexities of RKHSs become inhomogeneous. As mentioned above, it has been often reported that ℓ_1-MKL is outperformed by dense type MKL such as $\ell_{\frac{4}{3}}$-MKL in numerical experiments (Cortes et al., 2010). Our theoretical analysis in this section well support these experimental results.

6. Numerical comparison between homogeneous and inhomogeneous settings

Here we investigate numerically how the inhomogeneity of the complexities affects the performances using synthetic data. In particular, we numerically compare two situations: (a) all complexities of RKHSs are same (homogeneous situation) and (b) one RKHS is complex and other RKHSs are evenly simple (inhomogeneous situation).

The experimental settings are as follows. The input random variable is 20 dimensional vector $x = (x^{(1)}, \ldots, x^{(20)})$ where each element $x^{(m)}$ is independently identically distributed from the uniform distribution on $[0, 1]$:

$$x^{(m)} \sim \mathrm{Unif}([0,1]) \quad (m = 1, \ldots, 20).$$

For each coordinate $m = 1, \ldots, 20$, we put one Gaussian RKHS $\mathcal{H}_m$ with a Gaussian width ζ_m: the number of kernels is 20 ($M = 20$) and

$$k_m(x, x') = \exp\left(-\frac{(x^{(m)} - x'^{(m)})^2}{2\zeta_m^2}\right) \quad (m = 1, \ldots, 20),$$

for $x = (x^{(1)}, \ldots, x^{(20)})$ and $x' = (x'^{(1)}, \ldots, x'^{(20)})$. To generate the ground truth f^*, we randomly generated 5 center points $\mu_{i,m}$ $(i = 1, \ldots, 5)$ for each coordinate $m = 1, \ldots, 20$ where $\mu_{i,m}$ is independently generated by the uniform distribution on $[0, 1]$. Then we obtain the following form of the true function:

$$f^*(x) = \sum_{m=1}^{20} f_m^*(x),$$
$$\text{where} \quad f_m^*(x) = \sum_{i=1}^{5} \alpha_{i,m} \exp\left(-\frac{(x^{(m)} - \mu_{i,m})^2}{2\zeta_m^2}\right) \in \mathcal{H}_m,$$

for $x = (x_1, \ldots, x_m)$. Each coefficient $\alpha_{i,m}$ is independently identically distributed from the standard normal distribution. The output y is contaminated by a noise ϵ where the noise ϵ is distributed from the Gaussian distribution with mean 0 and standard deviation 0.1:

$$y = f_m^*(x) + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, 0.1).$$

We generated 200 or 400 realizations $\{(x_i, y_i)\}_{i=1}^n$ ($n = 200$ or $n = 400$), and estimated f^* using ℓ_p-MKL with $p = 1, 1.1, 1.2, \ldots, 3$ [4]. The estimator is computed with various regularization parameters $\lambda_1^{(n)}$. The generalization error $\|\hat{f} - f^*\|_{L_2(\Pi)}^2$ was numerically calculated. We repeated the experiments for 100 times, averaged the generalization errors over 100 repetitions for each p and each regularization parameter, and obtained the optimal average generalization error among all regularization parameters for each p. The true function was randomly generated for each repetition. We investigated the generalization errors in the following homogeneous and inhomogeneous settings:

1. (homogeneous) $\zeta_m = 0.5$ for $m = 1, \ldots, 20$.
2. (inhomogeneous) $\zeta_1 = 0.01$ and $\zeta_m = 0.5$ for $m = 2, \ldots, 20$.

The difference between the above homogeneous and inhomogeneous settings is the value of ζ_1; whether $\zeta_1 = 0.5$ or $\zeta_1 = 0.01$. The inhomogeneous situation is analogous to that investigated in Sec.5 where we assumed one RKHS is complex and the other RKHSs are evenly simple (small ζ_1 corresponds to a complex RKHS).

Figure 2 shows the average generalization errors in the homogeneous setting with (a) $n = 200$ and (c) $n = 400$, and the inhomogeneous setting with (b) $n = 200$ and (d) $n = 400$. Each broken line corresponds to one regularization parameter. The bold solid line shows the best (average) generalization error among all the regularization parameters. We can see that in the homogeneous setting ℓ_1-regularization shows the best performance, on the other hand, in the inhomogeneous setting the best performance is achieved at $p > 1$ for both $n = 200$ and 400. This experimental results beautifully matches the theoretical investigations.

7. Generalization of loss function

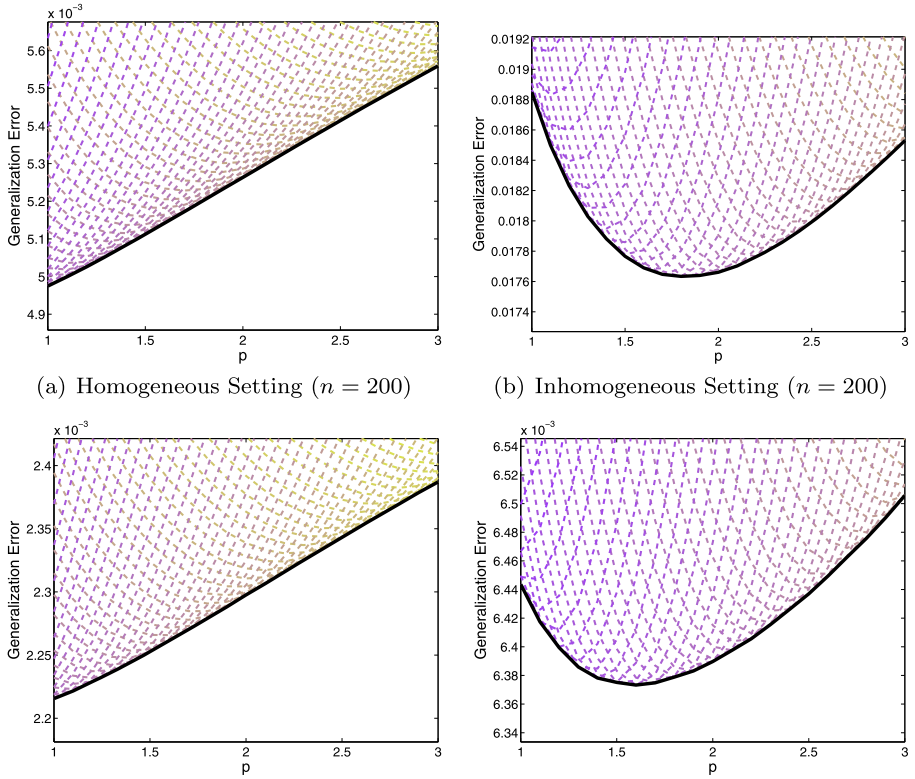
Here we discuss how a general loss function other than squared loss can be involved into our analysis. As in the standard local Rademacher complexity argument (Bartlett et al., 2005), we consider a class of loss functions that are Lipschitz continuous and strongly convex. Suppose that the loss function $\Psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfies Lipschitz continuity: for all $R > 0$, there exists a constant $T(R)$ such that

$$|\Psi(y, f_1) - \Psi(y, f_2)| \leq T(R)|f_1 - f_2| \\ (\forall f_1, f_2 \in \mathbb{R} \text{ such that } |f_1|, |f_2| \leq R, \ \forall y \in \mathbb{R}). \tag{21}$$

Moreover, suppose that, for all $y \in \mathbb{R}$, $\Psi(y, f)$ is a strongly convex with a modulus $\rho(R) > 0$:

$$\frac{\Psi(y, f_1) + \Psi(y, f_2)}{2} \geq \Psi\left(y, \frac{f_1 + f_2}{2}\right) + \frac{\rho(R)}{2}|f_1 - f_2|^2$$

[4] We included a bias term in this experiment, that is, we fitted $\hat{f}(x) + b$ to the data: $\min_{f_m, b} \frac{1}{n}\sum_{i=1}^n (y_i - \sum_{m=1}^M f_m(x_i) - b)^2 + \lambda_1^{(n)}\|f\|_{\ell_p}^2$.



(a) Homogeneous Setting ($n = 200$) (b) Inhomogeneous Setting ($n = 200$)

(c) Homogeneous Setting ($n = 400$) (d) Inhomogeneous Setting ($n = 400$)

FIG 2. *The expected generalization error* $\mathrm{E}[\|\hat{f} - f^*\|^2_{L_2(\Pi)}]$ *against the parameter* p *for* ℓ_p*-MKL. Each broken line corresponds to one regularization parameter. The bold solid line shows the best generalization error among all the regularization parameters.*

$$(\forall f_1, f_2 \in \mathbb{R} \text{ such that } |f_1|, |f_2| \leq R). \tag{22}$$

Some detailed discussions about these conditions and examples can be found in Bartlett et al. (2006). Under the loss functions satisfying these properties, we obtain simplified bound where some conditions can be omitted as follows:

- We can remove the condition $\frac{4\phi\sqrt{n}}{\kappa_M} \max\{\alpha_1^2, \beta_1^2, \frac{M\log(M)}{n}\}\eta(t') \leq \frac{1}{12}$,
- The term $\exp(-t')$ is not needed in the tail probability.

To obtain a fast convergence rate on a general loss functions Ψ, we move the regularization term in Eq. (1) into a constraint, and then consider the following optimization problem:

$$\hat{f} = \sum_{m=1}^{M} \hat{f}_m = \underset{\substack{f_m \in \mathcal{H}_m \ (m=1,\ldots,M), \\ \|f\|_\psi \leq \hat{R}}}{\arg\min} \frac{1}{n} \sum_{i=1}^{N} \Psi\left(y_i, \sum_{m=1}^{M} f_m(x_i)\right), \tag{23}$$

where $\hat{R}$ is a regularization parameter. The above optimization problem is essentially equivalent to the original formulation (1), but by considering the constraint type regularization instead of the penalty type regularization the theoretical analysis of statistical performance can be simplified.

We define Pg as the expectation of a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$:

$$Pg := \mathrm{E}_{(X,Y)\sim P}[g(X, Y)].$$

For notational simplicity, we write $P\Psi(f) = P\Psi(Y, f) = \mathrm{E}_{(X,Y)\sim P}[\Psi(Y, f(X))]$ for a function f. We suppose there exists a minimizer for $P\Psi(f)$ as follows.

Assumption 7. (Minimizer Existence Assumption)
There exists unique $f^* = (f_1^*, \ldots, f_M^*) \in \mathcal{H}^{\oplus M}$ *such that*

$$f^* = \sum_{m=1}^{M} f_m^* = \underset{f_m \in \mathcal{H}_m \ (m=1,\ldots,M)}{\arg\min} P\Psi\left(\sum_{m=1}^{M} f_m(X)\right). \tag{A7}$$

Note that, due to the incoherence assumption (Assumption 4) and the strong convexity (22) of the loss function, if there exists a minimizer, then that is automatically unique.

To bound the convergence rate on a general loss function, it is convenient to utilize *local Rademacher complexity* on ψ-norm ball. Let $\mathcal{H}_\psi^{(r)}(R) := \{f \in \mathcal{H}^{\oplus M} \mid \|f\|_{L_2(\Pi)} \le r,\ \|f\|_\psi \le R\}$. Then the local Rademacher complexity of $\mathcal{H}_\psi^{(r)}(R)$ is defined as

$$R_n(\mathcal{H}_\psi^{(r)}(R)) := \mathrm{E}_{\{\sigma_i, x_i\}_{i=1}^n}\left[\sup_{f \in \mathcal{H}_\psi^{(r)}(R)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right],$$

where $\sigma_i \in \{\pm 1\}$ is the i.i.d. Rademacher random variable with $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$. Evaluating the local Rademacher complexity is a key ingredient to show a fast convergence rate on a general loss function. We obtain the following estimation of the local Rademacher complexity (the proof will be given in Appendix F.4).

Lemma 7. *Let* $\{r_m\}_{m=1}^M$ *be arbitrary positive reals. Under Assumptions 2-5, there exists a constant* $\tilde{\phi}$ *depending on* $\{s_m\}_{m=1}^M, c, C_1$ *such that for all* n *satisfying* $\frac{\log(M)}{\sqrt{n}} \le 1$ *we have*

$$R_n(\mathcal{H}_\psi^{(r)}(R)) \le \tilde{\phi}\left(\alpha_1 \frac{r}{\sqrt{\kappa_M}} + \alpha_2 R + \beta_1 \frac{r}{\sqrt{\kappa_M}} + \beta_2 R + \sqrt{\frac{M\log(M)}{n}}\frac{r}{\sqrt{\kappa_M}}\right).$$

Finally note that the supremum norm of f with $\|f\|_\psi \le \hat{R}$ can be bounded as

$$\|f\|_\infty \le \sum_{m=1}^{M} \|f_m\|_\infty \le \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m} \le \|\mathbf{1}\|_{\psi^*}\|f\|_\psi \le \|\mathbf{1}\|_{\psi^*}\hat{R}.$$

Then, we obtain the *excess risk* bound as in the following theorem.

Theorem 8. *Suppose Assumptions 2-5 and 7 are satisfied and the loss function* Ψ *satisfies the conditions* (21) *and* (22). *Let* $\{r_m\}_{m=1}^M$ *be arbitrary positive reals that can depend on* n *and let* $\bar{T} = T(\|\mathbf{1}\|_{\psi^*}\hat{R})$ *and* $\bar{\rho} = \rho(\|\mathbf{1}\|_{\psi^*}\hat{R})$. *Set* $\hat{R} = \|f^*\|_\psi$. *Then there exists a constant* $\tilde{\phi}'$ *depending on* $\{s_m\}_{m=1}^M, c, C_1$ *such that for all* n *satisfying* $\frac{\log(M)}{\sqrt{n}} \leq 1$, *we have*

$$
\begin{aligned}
&P(\Psi(\hat{f}) - \Psi(f^*_{\hat{R}})) \\
\leq &\frac{\tilde{\phi}'\bar{\rho}}{\kappa_M}\left(\alpha_1^2 + \beta_1^2 + \frac{M\log(M)}{n}\right) + \tilde{\phi}'\frac{\bar{T}^2}{\bar{\rho}}\left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2 \\
&+ \frac{\{22\bar{T}\|\mathbf{1}\|_{\psi^*}\hat{R} + 27\bar{\rho}\}t}{n}, \qquad (24)
\end{aligned}
$$

with probability $1 - \exp(-t)$.

This can be shown by applying the bound of the local Rademacher complexity (Lemma 7) to Corollary 5.3 of Bartlett et al. (2005)[5]. Compared with the bound in Eq. (6), we notice that there is no $\exp(-t')$ term in the tail probability bound, and thus we don't need the condition $\frac{4\phi\sqrt{n}}{\kappa_M}\max\{\alpha_1^2, \beta_1^2, \frac{M\log(M)}{n}\}\eta(t') \leq \frac{1}{12}$. Because of this, the range of n where the error bound holds is relaxed compared with that in Theorem 1. These simplifications are due to the Lipschitz continuity of the loss function. In Theorem 1, we should have bounded the discrepancy between the empirical and population means of the squared loss: $\frac{1}{n}\sum_{i=1}^n(\hat{f}(x_i) - f^*(x_i))^2 - P(\hat{f} - f^*)^2$. Since the squared loss is not Lipschitz continuous, we required an additional bound for that discrepancy using Assumption 5 for the supremum norm, and it was shown that that discrepancy is negligible at the cost of $\exp(-t')$ in the tail probability. On the other hand, for Lipschitz continuous losses, we no longer need to bound such a quantity. Thus the tail probability loss $\exp(-t')$ is not induced.

Since the bound (24) is basically same as Eq.(6), we obtain the same discussions as in the previous sections. For example, in the homogeneous setting, we obtain the following convergence bound.

Lemma 9. *When* $s_m = s$ $(\forall m)$ *with some* $0 < s < 1$, *if we set* $\hat{R} = \|f^*\|_\psi$, *then for all* n *satisfying* $\frac{\log(M)}{\sqrt{n}} \leq 1$ *and* $n \geq (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi/M)^{\frac{4s}{1-s}}$, *and for all* $t \geq 1$, *we have*

$$
P(\Psi(\hat{f}) - \Psi(f^*)) \leq C\left\{M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} + \frac{M\log(M)}{n} + \frac{t}{n}\right\},
$$

with probability $1 - \exp(-t)$ *where* C *is a constant depending on* $\tilde{\phi}'$, κ_M, $\rho(\|\mathbf{1}\|_{\psi^*}\hat{R})$, *and* $T(\|\mathbf{1}\|_{\psi^*}\hat{R})$.

[5] In Corollary 5.3 of Bartlett et al. (2005), the range of the function class is assumed to be included in the interval $[-1, 1]$. Here we utilize more general settings where the interval is $[-a, a]$ and $\|\mathbf{1}\|_{\psi^*}\hat{R}$ is substituted to a. See Lemma 9 of Kloft and Blanchard (2011).

8. Conclusion and future work

We have shown a unifying framework to derive the learning rate of MKL with arbitrary mixed-norm-type regularization. To analyze the general result, we considered two situations: homogeneous settings and inhomogeneous settings. We have seen that the convergence rate of ℓ_p-MKL obtained in homogeneous settings is tighter and requires less restrictive condition than existing results. We have also shown convergence rates of some examples (elasticnet-MKL and VSKL), and proved the derived learning rate is minimax optimal when ψ-norm is isotropic. An interesting consequence was that ℓ_1-regularization is optimal among all isotropic ψ-norm regularization in homogeneous settings. In the analysis of inhomogeneous settings, we have shown that the dense type regularization can outperform the sparse ℓ_1-regularization using analytically obtained bounds and numerically computed bounds. We observed that our bound well explains the experimental results favorable for dense type MKL. Finally we numerically investigated the generalization errors of ℓ_p-MKL in a homogeneous setting and an inhomogeneous setting. The numerical experiments supported the theoretical findings that ℓ_1-regularization is optimal in homogeneous settings but, on the other hand, dense type regularizations are preferred in inhomogeneous settings. This is the first result that suggests that the inhomogeneity of the complexities of RKHSs well justifies the favorable performances for dense type MKL.

An interesting future work is about the $\frac{M\log(M)}{n}$ term appeared in the bound Eq. (8). Because of this term, our bound is $O(M\log(M))$ with respect to M while in the existing work that is $O(\sqrt{\log(M)} \vee M^{1-\frac{1}{p}})$ for ℓ_p-MKL. Therefore our bound is not tight in the global bound regime ($n \leq M^{\frac{2}{p}} R_p^{-2} \log(M)^{\frac{1+s}{s}}$ for ℓ_p-MKL). It is an interesting issue to clarify whether the term $\frac{M\log(M)}{n}$ can be replaced by other tighter bounds or not. To do so, it might be helpful to combine our technique developed in this paper and that developed by Kloft and Blanchard (2011) where the local Rademacher complexity for ℓ_p-MKL is derived.

Appendix A: Relation between entropy number and spectral condition

Associated with the ϵ-covering number, the *i-th entropy number* $e_i(\mathcal{H}_m \to L_2(\Pi))$ is defined as the infimum over all $\varepsilon > 0$ for which $N(\varepsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi)) \leq 2^{i-1}$. If the spectral assumption (A3) for $0 < s < 1$ (Assumption 3) and the boundedness assumption (A2) hold, the relation (2) implies that the i-th entropy number is bounded as

$$e_i(\mathcal{H}_m \to L_2(\Pi)) \leq C i^{-\frac{1}{2s}}, \tag{25}$$

where C is a constant. To bound empirical process a bound of the entropy number with respect to the empirical distribution is needed. The following proposition gives an upper bound of that (see Corollary 7.31 of Steinwart (2008), for example).

Proposition 10. *If there exists constants $0 < s < 1$ and $C \geq 1$ such that $e_i(\mathcal{H}_m \to L_2(\Pi)) \leq C i^{-\frac{1}{2s}}$, then there exists a constant $c_s > 0$ only depending on s such that*

$$\mathrm{E}_{D_n \sim \Pi^n}[e_i(\mathcal{H}_m \to L_2(D_n))] \leq c_s C (\min(i, n))^{\frac{1}{2s}} i^{-\frac{1}{s}},$$

in particular $\mathrm{E}_{D_n \sim \Pi^n}[e_i(\mathcal{H}_m \to L_2(D_n))] \leq c_s C i^{-\frac{1}{2s}}$.

Appendix B: Basic propositions

The following two propositions are keys to prove Theorem 1. Let $(\sigma_i)_{i=1}^n$ be i.i.d. Rademacher random variables, i.e., $\sigma_i \in \{\pm 1\}$ and $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.

Proposition 11. (Steinwart, 2008, Theorem 7.16) *Let $\mathcal{B}_{\sigma,a,b} \subset \mathcal{H}_m$ be a set such that $\mathcal{B}_{B,a,b} = \{f_m \in \mathcal{H}_m \mid \|f_m\|_{L_2(\Pi)} \leq B, \|f_m\|_{\mathcal{H}_m} \leq a, \|f_m\|_\infty \leq b\}$. Assume that there exist constants $0 < s < 1$ and $0 < \tilde{c}_s$ such that*

$$\mathrm{E}_{D_n}[e_i(\mathcal{H}_m \to L_2(D_n))] \leq \tilde{c}_s i^{-\frac{1}{2s}}.$$

Suppose that $\{\xi_i\}_{i=1}^n$ is a sequence of i.i.d. sub-Gaussian random variables satisfying $\mathrm{E}[e^{\xi_i t}] \leq \exp(t^2/2)$ for all $t \in \mathbb{R}$. Then there exists a constant C'_s depending only s such that

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{B}_{B,a,b}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i f_m(x_i)\right|\right] \leq C'_s \left(\frac{B^{1-s}(\tilde{c}_s a)^s}{\sqrt{n}} \vee (\tilde{c}_s a)^{\frac{2s}{1+s}} b^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}}\right). \tag{26}$$

In particular, this bound hods for the Rademacher random variable $\xi_i = \sigma_i$ and the standard normal random variable $\xi_i \sim N(0,1)$.

Moreover, for $s = 0$, the following proposition holds.

Proposition 12. *Under the same notation in Proposition 11 and the spectral assumption (Assumption 3) for $\mathcal{H}_m$ with $s_m = 0$, it holds that*

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{B}_{\sigma,a,b}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i f_m(x_i)\right|\right] \leq \sqrt{\frac{c}{n}} B.$$

Proof. Since the RKHS $\mathcal{H}_m$ is finite dimensional (say d-dimensional), there exists $\varphi : \mathcal{X} \to \mathbb{R}^d$ such that each element f can be expressed by $f(x) = \beta^\top \varphi(x)$ by a vector $\beta \in \mathbb{R}^d$. Let $\Sigma = \mathrm{E}[\varphi(X)\varphi(X)^\top] \in \mathbb{R}^{d \times d}$. Then, since the dimension of the RKHS $\mathcal{H}_m$ is d, we have $\Sigma \succ O$. Thus,

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{B}_{B,a,b}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i f_m(x_i)\right|\right]$$

$$
\begin{aligned}
&= \mathrm{E}\left[\sup_{\beta_m : f_m = \beta_m^\top \varphi \in \mathcal{B}_{B,a,b}} \left| \frac{1}{n}\sum_{i=1}^n \xi_i \beta_m^\top \varphi(x_i) \right| \right] \\
&\leq \mathrm{E}\left[\sup_{\beta_m : f_m = \beta_m^\top \varphi \in \mathcal{B}_{B,a,b}} \|\beta_m\|_\Sigma \left\| \frac{1}{n}\sum_{i=1}^n \xi_i \varphi(x_i) \right\|_{\Sigma^{-1}} \right] \\
&\leq \mathrm{E}\left[B \left\| \frac{1}{n}\sum_{i=1}^n \xi_i \varphi(x_i) \right\|_{\Sigma^{-1}} \right] \\
&\leq B \sqrt{\mathrm{E}\left[\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \xi_i \xi_j \varphi(x_i)^\top \Sigma^{-1} \varphi(x_j) \right]} \\
&= B\sqrt{\frac{1}{n}\mathrm{Tr}[\Sigma^{-1}\Sigma]} \leq \sqrt{\frac{d}{n}} B \leq \sqrt{\frac{c}{n}} B,
\end{aligned}
$$

where the last inequality is by Assumption 3. □

Proposition 13. (Talagrand's Concentration Inequality (Talagrand (1996); Bousquet (2002))) *Let $\mathcal{G}$ be a function class on $\mathcal{X}$ that is separable with respect to ∞-norm, and $\{x_i\}_{i=1}^n$ be i.i.d. random variables with values in $\mathcal{X}$. Furthermore, let $B \geq 0$ and $U \geq 0$ be $B := \sup_{g\in\mathcal{G}} \mathrm{E}[(g - \mathrm{E}[g])^2]$ and $U := \sup_{g\in\mathcal{G}} \|g\|_\infty$, then there exists a universal constant K such that, for $Z := \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n g(x_i) - \mathrm{E}[g]\right|$, we have*

$$
P\left(Z \geq K \left[\mathrm{E}[Z] + \sqrt{\frac{Bt}{n}} + \frac{Ut}{n} \right] \right) \leq e^{-t}.
$$

Since the above proposition assumes that the functions are uniformly bounded, it can not be applied to evaluate the supremum of an empirical process with unbounded summands; in particular, $\sup_{g\in\mathcal{G}} |\sum_{i=1}^n \epsilon_i g(x_i)|$ where ϵ_i is an i.i.d. Gaussian noise. To bound this type of Gaussian process, we utilize the Gaussian concentration inequality.

Proposition 14 (Gaussian concentration inequality (Theorem 2.5.8 in Giné and Nickl (2015)))**.** *Let $\{\xi_i\}_{i=1}^n$ be i.i.d. Gaussian sequence with mean 0 and variance L^2, and $\{x_i\}_{i=1}^n \subset \mathcal{X}$ be a given set of input variables. Then, for a set $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ which is separable with respect to L_∞-norm and $\sup_{f\in\mathcal{G}} \left|\sum_{i=1}^n \frac{1}{n}\xi_i g(x_i)\right| < \infty$ almost surely, it holds that for every $r > 0$,*

$$
\begin{aligned}
P\Bigg(\sup_{g\in\mathcal{G}} \left| \sum_{i=1}^n \frac{1}{n}\xi_i g(x_i) \right| \geq \mathrm{E}_\xi \left[\sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^n \xi_i g(x_i) \right| \,\middle|\, \{x_i\}_{i=1}^n \right] \\
+ \frac{\sqrt{2}L\|\mathcal{G}\|_n}{\sqrt{n}}\sqrt{r} \,\middle|\, \{x_i\}_{i=1}^n \Bigg) \leq e^{-r}
\end{aligned}
$$

where $\|\mathcal{G}\|_n^2 = \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(x_i)^2$. Here the probability is taken with respect to $\{\xi_i\}_{i=1}^n$.

Here, the bound still depends on the observations $\{x_i\}_{i=1}^n$. To derive the population bound, we utilize the following bound for the *self-bounding random variable*.

Definition 15 (Self-bounding random variable)**.** *Let $X_1, \dots, X_n$ be a sequence of independent random variables. A random variable $Z = f(X_1, \dots, X_n)$ is called self-bounding if there exists a sequence of random variables $Z_k = f_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)$ $(k = 1, \dots, n)$ such that*

$$0 \leq Z - Z_k \leq 1 \quad (1 \leq k \leq n), \quad \sum_{k=1}^{n} (Z - Z_k) \leq Z.$$

Proposition 16 (Theorem 3.3.15 of Giné and Nickl (2015))**.** *Let Z be a self-bounding random variable. Then, for $t > 0$, it holds that*

$$P(Z \geq \mathrm{E}[Z] + \sqrt{\mathrm{E}[Z]t} + t/3) \leq e^{-t}.$$

In particular, we have

$$P(Z \geq 2\mathrm{E}[Z] + t) \leq e^{-t}.$$

(i) For example, $Z = \sup_{g \in \mathcal{G}} \sum_{i=1}^n g(X_i)$ is a self-bounding random variable, when $\mathcal{G}$ is separable with respect to L_∞-norm and $0 \leq g(x) \leq 1$ for all $x \in \mathcal{X}$ and $g \in \mathcal{G}$.

(ii) $Z = f(x_1, \dots, x_n) = \mathrm{E}_\xi[\sup_{g \in \mathcal{G}} |\sum_{i=1}^n \xi_i g(x_i)| \ |\{x_i\}_{i=1}^n]$ is also self-bounding, when $\mathcal{G}$ is separable with respect to L_∞-norm and $0 \leq g(x) \leq 1$ for all $x \in \mathcal{X}$ and $g \in \mathcal{G}$ and ξ_i is independent standard normal. This can be checked as $Z_k = \mathrm{E}_\xi[\sup_{g \in \mathcal{G}} |\sum_{i=1, i \neq k}^n \xi_i g(x_i)| \ |\{x_i\}_{i=1}^n]$ and $Z \leq \mathrm{E}_\xi[\sup_{g \in \mathcal{G}} |\sum_{i=1, i \neq k}^n \xi_i g(x_i)| \ |\{x_i\}_{i=1}^n] + \mathrm{E}_\xi[\sup_{g \in \mathcal{G}} |\xi_k g(x_k)| | x_k] \leq Z_k + \mathrm{E}_\xi[|\xi_k|] \leq Z_k + \sqrt{\mathrm{E}_\xi[|\xi_k|^2]} \leq Z_k + 1$, and the convexity of $(\xi_1, \dots, \xi_n) \mapsto \sup_{g \in \mathcal{G}} |\sum_{i=1}^n \xi_i g(x_i)|$ gives $Z \geq \mathrm{E}_{(\xi)_{i \neq k}}[\sup_{g \in \mathcal{G}} |\mathrm{E}_{\xi_k}[\sum_{i=1}^n \xi_i g(x_i)]| \ |\{x_i\}_{i=1}^n] = Z_k$.

Therefore, with Proposition 16, we obtain the following lemma.

Lemma 17. *Let $\{\xi_i\}_{i=1}^n$ be i.i.d. Gaussian sequence with mean 0 and variance L^2. For a set $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ which is separable with respect to L_∞-norm and $\sup_{g \in \mathcal{G}} \|g\|_\infty < U$, it holds that for every $t > 0$,*

$$P(\|\mathcal{G}\|_n^2 \geq 2\|\mathcal{G}\|_{L_2(\Pi)}^2 + U^2 t/n) \leq e^{-t},$$

and

$$P\left(\mathrm{E}_\xi\left[\sup_{g \in \mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i g(x_i)\right| \middle| \{x_i\}_{i=1}^n\right] \geq 2\mathrm{E}_\xi\left[\sup_{g \in \mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i g(x_i)\right|\right] + \frac{LUt}{n}\right) \leq e^{-t}.$$

As a consequence, we have the following lemma which corresponds to a Gaussian process version of Proposition 13.

Lemma 18. *Under the same assumptions with Proposition 13 except that* $B := \sup_{g\in\mathcal{G}} \mathrm{E}[g^2]$, $U := \sup_{g\in\mathcal{G}} \|g\|_\infty$ *and* $Z := \sup_{g\in\mathcal{G}} |\sum_{i=1}^n \frac{1}{n}\xi_i g(x_i)|$ *where* $\{\xi_i\}_{i=1}^n$ *is a sequence of i.i.d. normal variables with mean* 0 *and variance* L^2. *Then, there exists a universal constant* $K > 0$ *such that*

$$P_{x,\xi}\left(Z \geq K\left[\mathrm{E}[Z] + L\sqrt{\frac{Bt}{n}} + \frac{LUt}{n}\right]\right) \leq 3e^{-t},$$

where the probability is taken with respect to $\{\xi_i\}_{i=1}^n$ *and* $\{x_i\}_{i=1}^n$.

Proof. Combining Proposition 14 and Lemma 17, it holds that, for $t > 0$,

$$P\left(Z \geq 2\mathrm{E}\left[Z\right] + \frac{LUt}{n} + \frac{\sqrt{2}L\sqrt{2\|\mathcal{G}\|_{L_2(\Pi)}^2 + U^2t/n}}{\sqrt{n}}\sqrt{t} + \frac{LUt}{n}\right)$$
$$\leq e^{-t} + e^{-t} + e^{-t}.$$

Then, by arranging the terms in the left hand side, we obtain the assertion. □

Appendix C: Proof of Theorem 1

Let $r_m > 0$ $(m = 1, \ldots, M)$ be arbitrary positive reals. Given $\{r_m\}_{m=1}^M$, we determine $U_{n,s_m}^{(m)}(f_m)$ as follows:

$$U_{n,s_m}^{(m)}(f_m) := 3\left(\frac{r_m^{-s_m}}{\sqrt{n}} \vee \frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\right)\left(\|f_m\|_{L_2(\Pi)} + s_m r_m \|f_m\|_{\mathcal{H}_m}\right)$$
$$+ \sqrt{\frac{\log(M)}{n}}\|f_m\|_{L_2(\Pi)}.$$

It is easy to see $U_{n,s_m}^{(m)}(f_m)$ is an upper bound of the quantity $\frac{\|f_m\|_{L_2(\Pi)}^{1-s_m}\|f_m\|_{\mathcal{H}_m}^{s_m}}{\sqrt{n}} \vee \frac{\|f_m\|_{L_2(\Pi)}^{\frac{(1-s_m)^2}{1+s_m}}\|f_m\|_{\mathcal{H}_m}^{\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}$ (this corresponds to the RHS of Eq. (26)) because

$$\frac{\|f_m\|_{L_2(\Pi)}^{1-s_m}\|f_m\|_{\mathcal{H}_m}^{s_m}}{\sqrt{n}} = \frac{r_m^{1-s_m}}{\sqrt{n}}\left(\frac{\|f_m\|_{L_2(\Pi)}}{r_m}\right)^{1-s_m}\|f_m\|_{\mathcal{H}_m}^{s_m}$$
$$\overset{\text{(Young)}}{\leq} \frac{r_m^{1-s_m}}{\sqrt{n}}\left((1-s_m)\frac{\|f_m\|_{L_2(\Pi)}}{r_m} + s_m\|f_m\|_{\mathcal{H}_m}\right)$$
$$\leq \frac{r_m^{-s_m}}{\sqrt{n}}\left(\|f_m\|_{L_2(\Pi)} + s_m r_m\|f_m\|_{\mathcal{H}_m}\right), \tag{27}$$

where we used Young's inequality $a^{1-s_m}b^{s_m} \leq (1-s_m)a + s_m b$ in the second line, and similarly we obtain

$$\frac{\|f_m\|_{L_2(\Pi)}^{\frac{(1-s_m)^2}{1+s_m}}\|f_m\|_{\mathcal{H}_m}^{\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}$$

$$\leq \frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}} \left(\|f_m\|_{L_2(\Pi)} + \frac{s_m(3-s_m)}{1+s_m} r_m \|f_m\|_{\mathcal{H}_m} \right)$$
$$\leq 3 \frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}} \left(\|f_m\|_{L_2(\Pi)} + s_m r_m \|f_m\|_{\mathcal{H}_m} \right),$$

where we used $\frac{s_m(3-s_m)}{1+s_m} \leq 3s_m$ in the last inequality.

Now we define

$$\phi := \max\left(KL\left[2\tilde{C}_* + 1 + C_1\right], K\left[2C_1\tilde{C}_* + C_1 + C_1^2\right] \right),$$

where $\tilde{C}^*$ is a constant defined later in Lemma 23, C_1 is the one introduced in Assumption 5, K is the maximum of the universal constants appeared in Talagrand's concentration inequality (Proposition 13) and Lemma 18, and L is the one introduced in Assumption 1 to bound the magnitude of noise. Remind the definition of $\eta(t)$:

$$\eta(t) := \eta_n(t) = \max(1, \sqrt{t}, t/\sqrt{n}).$$

We define events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(t')$ as

$$\mathscr{E}_1(t) = \left\{ \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f_m(x_i) \right| \leq \phi U_{n,s_m}^{(m)}(f_m) \eta(t), \ \forall f_m \in \mathcal{H}_m \ (m = 1, \ldots, M) \right\}, \tag{28}$$

$$\mathscr{E}_2(t') = \left\{ \left| \left\| \sum_{m=1}^M f_m \right\|_n^2 - \left\| \sum_{m=1}^M f_m \right\|_{L_2(\Pi)}^2 \right| \leq \phi\sqrt{n} \left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m) \right)^2 \eta(t'), \right.$$
$$\left. \forall f_m \in \mathcal{H}_m \ (m = 1, \ldots, M) \right\}. \tag{29}$$

Using Lemmas 24 and 25 that will be shown in Appendix D, we see that the events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(t')$ occur with probability no less than $1 - \exp(-t)$ and $1 - \exp(-t')$ respectively as in the following Lemma.

Lemma 19. *Under the Basic Assumption (Assumption 1), the Spectral Assumption (Assumption 3) and the Embedded Assumption (Assumption 5), the probabilities of $\mathscr{E}_1(t)$ and $\mathscr{E}_2$ are bounded as*

$$P(\mathscr{E}_1(t)) \geq 1 - \exp(-t), \quad P(\mathscr{E}_2(t')) \geq 1 - \exp(-t').$$

Proof. Lemma 25 immediately gives $P(\mathscr{E}_1(t)) \geq 1 - \exp(-t)$ by noticing $\bar{\phi}$ in the statement of Lemma 25 satisfies $\bar{\phi} \leq \phi$. Moreover, since $\bar{\phi}'$ in the statement of Lemma 24 satisfies $\bar{\phi}' \leq \phi$, we have $P(\mathscr{E}_2(t')) \geq 1 - \exp(-t')$ by Lemma 24. □

Remind the definition (4) of $\alpha_1, \alpha_2, \beta_1, \beta_2$:

$$\alpha_1 = 3 \left(\sum_{m=1}^M \frac{r_m^{-2s_m}}{n} \right)^{\frac{1}{2}}, \quad \alpha_2 = 3 \left\| \left(\frac{s_m r_m^{1-s_m}}{\sqrt{n}} \right)_{m=1}^M \right\|_{\psi^*},$$

$$\beta_1 = 3\left(\sum_{m=1}^{M} \frac{r_m^{-\frac{2s_m(3-s_m)}{1+s_m}}}{n^{\frac{2}{1+s_m}}}\right)^{\frac{1}{2}}, \quad \beta_2 = 3\left\|\left(\frac{s_m r_m^{\frac{(1-s_m)^2}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\right)_{m=1}^{M}\right\|_{\psi^*}, \tag{30}$$

for given reals $\{r_m\}_{m=1}^M$. The following theorem immediately gives Theorem 1.

Theorem 20. *Suppose Assumptions 1-4 are satisfied. Let $\{r_m\}_{m=1}^M$ be arbitrary positive reals that can depend on n, and assume $\lambda_1^{(n)} \geq \left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2$. Then for all n and t' that satisfy $\frac{\log(M)}{\sqrt{n}} \leq 1$ and $\frac{4\phi\sqrt{n}}{\kappa_M}\max\{\alpha_1^2, \beta_1^2, \frac{M\log(M)}{n}\}\eta(t') \leq \frac{1}{12}$ and for all $t \geq 1$, we have*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{24\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2 + \beta_1^2 + \frac{M\log(M)}{n}\right) + 4\lambda_1^{(n)}\|f^*\|_\psi^2.$$

with probability $1 - \exp(-t) - \exp(-t')$.

Proof of Theorem 20. By the assumption of the theorem, we can assume Lemma 19 holds, that is, the event $\mathscr{E}_1(t) \cap \mathscr{E}_2(t')$ occurs with probability $1 - \exp(-t) - \exp(-t')$. Below we discuss on the event $\mathscr{E}_1(t) \cap \mathscr{E}_2(t')$.

Since $y_i = f^*(x_i) + \epsilon_i$, we have

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \lambda_1^{(n)}\|\hat{f}\|_\psi^2 \\
\leq&(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 - \|\hat{f} - f^*\|_n^2) + \frac{2}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) + \lambda_1^{(n)}\|f^*\|_\psi^2.
\end{aligned}$$

Here on the event $\mathscr{E}_2(t')$, the above inequality gives

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \lambda_1^{(n)}\|\hat{f}\|_\psi^2 \\
\leq&\phi\sqrt{n}\left(\sum_{m=1}^{M} U_{n,s_m}^{(m)}(\hat{f}_m - f_m^*)\right)^2 \eta(t') + \frac{2}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \\
&+ \lambda_1^{(n)}\|f^*\|_\psi^2.
\end{aligned} \tag{31}$$

Before we prove the statements, we show an upper bound of $\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)$ required in the proof. By definition, we have

$$\begin{aligned}
&U_{n,s_m}^{(m)}(f_m) \\
=&3\left(\frac{r_m^{-s_m}}{\sqrt{n}} \vee \frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\right)\left(\|f_m\|_{L_2(\Pi)} + s_m r_m\|f_m\|_{\mathcal{H}_m}\right) + \sqrt{\frac{\log(M)}{n}}\|f_m\|_{L_2(\Pi)} \\
\leq&3\frac{r_m^{-s_m}}{\sqrt{n}}\left(\|f_m\|_{L_2(\Pi)} + s_m r_m\|f_m\|_{\mathcal{H}_m}\right) \\
\leq&+3\frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\left(\|f_m\|_{L_2(\Pi)} + s_m r_m\|f_m\|_{\mathcal{H}_m}\right)
\end{aligned} \tag{32}$$

$$+\sqrt{\frac{\log(M)}{n}}\|f_m\|_{L_2(\Pi)}. \tag{33}$$

Now the sum of the first term is bounded as

$$\begin{aligned}
&\sum_{m=1}^{M} 3\frac{r_m^{-s_m}}{\sqrt{n}}\left(\|f_m\|_{L_2(\Pi)} + s_m r_m \|f_m\|_{\mathcal{H}_m}\right) \\
=&3\sum_{m=1}^{M}\frac{r_m^{-s_m}}{\sqrt{n}}\|f_m\|_{L_2(\Pi)} + 3\sum_{m=1}^{M}\frac{s_m r_m^{1-s_m}}{\sqrt{n}}\|f_m\|_{\mathcal{H}_m} \\
\leq&3\left(\sum_{m=1}^{M}\frac{r_m^{-2s_m}}{n}\right)^{\frac{1}{2}}\left(\sum_{m=1}^{M}\|f_m\|_{L_2(\Pi)}^2\right)^{\frac{1}{2}} + 3\left\|\left(\frac{s_m r_m^{1-s_m}}{\sqrt{n}}\right)_{m=1}^{M}\right\|_{\psi^*}\|f\|_{\psi},
\end{aligned}$$

where we used Cauchy-Schwarz inequality and the duality of the norm in the last inequality. The sum of the second term of the RHS of Eq. (33) is bounded as

$$\begin{aligned}
&\sum_{m=1}^{M} 3\frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\left(\|f_m\|_{L_2(\Pi)} + s_m r_m \|f_m\|_{\mathcal{H}_m}\right) \\
=&3\sum_{m=1}^{M}\frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\|f_m\|_{L_2(\Pi)} + 3\sum_{m=1}^{M}\frac{s_m r_m^{\frac{(1-s_m)^2}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\|f_m\|_{\mathcal{H}_m} \\
\leq&3\left(\sum_{m=1}^{M}\frac{r_m^{-\frac{2s_m(3-s_m)}{1+s_m}}}{n^{\frac{2}{1+s_m}}}\right)^{\frac{1}{2}}\left(\sum_{m=1}^{M}\|f_m\|_{L_2(\Pi)}^2\right)^{\frac{1}{2}} + 3\left\|\left(\frac{s_m r_m^{\frac{(1-s_m)^2}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\right)_{m=1}^{M}\right\|_{\psi^*}\|f\|_{\psi},
\end{aligned}$$

where we used Cauchy-Schwarz inequality and the duality of the norm in the last inequality. Finally we have the following bound of the third term of the RHS of Eq. (33):

$$\sum_{m=1}^{M}\sqrt{\frac{\log(M)}{n}}\|f_m\|_{L_2(\Pi)} \leq \sqrt{\frac{M\log(M)}{n}}\left(\sum_{m=1}^{M}\|f_m\|_{L_2(\Pi)}^2\right)^{\frac{1}{2}}.$$

Combine these inequalities and the relation $\sum_{m=1}^{M}\|f_m\|_{L_2(\Pi)}^2 \leq \frac{1}{\kappa_M}\|f\|_{L_2(\Pi)}^2$ (Assumption 4) to obtain

$$\begin{aligned}
&\sum_{m=1}^{M} U_{n,s_m}^{(m)}(f_m) \\
\leq&3\left(\sum_{m=1}^{M}\frac{r_m^{-2s_m}}{n}\right)^{\frac{1}{2}}\frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}} + 3\left\|\left(\frac{s_m r_m^{1-s_m}}{\sqrt{n}}\right)_{m=1}^{M}\right\|_{\psi^*}\|f\|_{\psi}
\end{aligned}$$

$$+3\left(\sum_{m=1}^{M}\frac{r_m^{-\frac{2s_m(3-s_m)}{1+s_m}}}{n^{\frac{2}{1+s_m}}}\right)^{\frac{1}{2}}\frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}+3\left\|\left(\frac{s_m r_m^{\frac{(1-s_m)^2}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\right)_{m=1}^{M}\right\|_{\psi^*}\|f\|_{\psi}$$
$$+\sqrt{\frac{M\log(M)}{n}}\frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}. \tag{34}$$

Then by the definition (4) of $\alpha_1,\alpha_2,\beta_1,\beta_2$, we have

$$\begin{aligned}&\sum_{m=1}^{M}U_{n,s_m}^{(m)}(f_m)\\ \le&\alpha_1\frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}+\alpha_2\|f\|_{\psi}+\beta_1\frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}+\beta_2\|f\|_{\psi}+\sqrt{\frac{M\log(M)}{n}}\frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}.\end{aligned} \tag{35}$$

Step 1.

By Eq. (35), the first term on the RHS of Eq. (31) can be upper bounded as

$$\begin{aligned}&\phi\sqrt{n}\left(\sum_{m=1}^{M}U_{n,s_m}^{(m)}(\hat{f}_m-f_m^*)\right)^2\eta(t')\\ \le&4\phi\sqrt{n}\Big(\alpha_1^2\frac{\|\hat{f}-f^*\|_{L_2(\Pi)}^2}{\kappa_M}+\alpha_2^2\|\hat{f}-f^*\|_{\psi}^2+\beta_1^2\frac{\|\hat{f}-f^*\|_{L_2(\Pi)}^2}{\kappa_M}+\\ &\qquad\beta_2^2\|\hat{f}-f^*\|_{\psi}^2+\frac{M\log(M)}{n}\frac{\|\hat{f}-f^*\|_{L_2(\Pi)}^2}{\kappa_M}\Big)\eta(t')\\ \le&\frac{4\phi\sqrt{n}}{\kappa_M}\alpha_1^2\eta(t')\left(\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left(\frac{\alpha_2}{\alpha_1}\right)^2\|\hat{f}-f^*\|_{\psi}^2\right)\\ &+\frac{4\phi\sqrt{n}}{\kappa_M}\beta_1^2\eta(t')\left(\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left(\frac{\beta_2}{\beta_1}\right)^2\|\hat{f}-f^*\|_{\psi}^2\right)\\ &+\frac{4\phi\sqrt{n}}{\kappa_M}\frac{M\log(M)}{n}\eta(t')\|\hat{f}-f^*\|_{L_2(\Pi)}^2.\end{aligned}$$

By assumption, we have $\frac{4\phi\sqrt{n}}{\kappa_M}\max\{\alpha_1^2,\beta_1^2,\frac{M\log(M)}{n}\}\eta(t')\le\frac{1}{12}$. Hence the RHS of the above inequality is bounded by

$$\begin{aligned}&\phi\sqrt{n}\left(\sum_{m=1}^{M}U_{n,s_m}^{(m)}(\hat{f}_m-f_m^*)\right)^2\eta(t')\\ \le&\frac{1}{4}\left\{\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left[\left(\frac{\alpha_2}{\alpha_1}\right)^2+\left(\frac{\beta_2}{\beta_1}\right)^2\right]\|\hat{f}-f^*\|_{\psi}^2\right\}.\end{aligned} \tag{36}$$

Step 2. On the event $\mathscr{E}_1(t)$, we have

$$
\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\sum_{m=1}^{M}\epsilon_i(\hat{f}_m(x_i)-f_m^*(x_i)) \le 2\sum_{m=1}^{M}\eta(t)\phi U_{n,s_m}^{(m)}(\hat{f}_m-f_m^*)\\
&\le 2\eta(t)\phi\left[\alpha_1\frac{\|\hat{f}-f^*\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}+\alpha_2\|\hat{f}-f^*\|_\psi+\beta_1\frac{\|\hat{f}-f^*\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}+\beta_2\|\hat{f}-f^*\|_\psi\right.\\
&\quad\left.+\sqrt{\frac{M\log(M)}{n}}\frac{\|\hat{f}-f^*\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}\right] \qquad (\because \text{Eq.(34)})\\
&\le 2\frac{\eta(t)\phi\alpha_1}{\sqrt{\kappa_M}}\left(\|\hat{f}-f^*\|_{L_2(\Pi)}+\frac{\alpha_2}{\alpha_1}\|\hat{f}-f^*\|_\psi\right)\\
&\quad+2\frac{\eta(t)\phi\beta_1}{\sqrt{\kappa_M}}\left(\|\hat{f}-f^*\|_{L_2(\Pi)}+\frac{\beta_2}{\beta_1}\|\hat{f}-f^*\|_\psi\right)\\
&\quad+2\frac{\eta(t)\phi}{\sqrt{\kappa_M}}\sqrt{\frac{M\log(M)}{n}}\|\hat{f}-f^*\|_{L_2(\Pi)}\\
&\le \frac{12\eta(t)^2\phi^2\alpha_1^2}{\kappa_M}+\frac{1}{24}\left(\|\hat{f}-f^*\|_{L_2(\Pi)}+\frac{\alpha_2}{\alpha_1}\|\hat{f}-f^*\|_\psi\right)^2\\
&\quad+\frac{12\eta(t)^2\phi^2\beta_1^2}{\kappa_M}+\frac{1}{24}\left(\|\hat{f}-f^*\|_{L_2(\Pi)}+\frac{\beta_2}{\beta_1}\|\hat{f}-f^*\|_\psi\right)^2\\
&\quad+\frac{6\eta(t)^2\phi^2}{\kappa_M}\frac{M\log(M)}{n}+\frac{1}{12}\|\hat{f}-f^*\|_{L_2(\Pi)}^2\\
&\le \frac{12\eta(t)^2\phi^2\alpha_1^2}{\kappa_M}+\frac{1}{12}\left[\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left(\frac{\alpha_2}{\alpha_1}\right)^2\|\hat{f}-f^*\|_\psi^2\right]\\
&\quad+\frac{12\eta(t)^2\phi^2\beta_1^2}{\kappa_M}+\frac{1}{12}\left[\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left(\frac{\beta_2}{\beta_1}\right)^2\|\hat{f}-f^*\|_\psi^2\right]\\
&\quad+\frac{6\eta(t)^2\phi^2}{\kappa_M}\frac{M\log(M)}{n}+\frac{1}{12}\|\hat{f}-f^*\|_{L_2(\Pi)}^2\\
&\le \frac{12\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2+\beta_1^2+\frac{M\log(M)}{n}\right)\\
&\quad+\frac{1}{4}\left\{\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left[\left(\frac{\alpha_2}{\alpha_1}\right)^2+\left(\frac{\beta_2}{\beta_1}\right)^2\right]\|\hat{f}-f^*\|_\psi^2\right\}. \qquad (37)
\end{aligned}
$$

Step 3.

Substituting the inequalities (36) and (37) to Eq. (31), we obtain

$$
\begin{aligned}
&\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\lambda_1^{(n)}\|\hat{f}\|_\psi^2\\
&\le\frac{12\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2+\beta_1^2+\frac{M\log(M)}{n}\right)
\end{aligned}
$$

$$+\frac{1}{2}\left\{\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\left[\left(\frac{\alpha_2}{\alpha_1}\right)^2+\left(\frac{\beta_2}{\beta_1}\right)^2\right]\|\hat{f}-f^*\|_\psi^2\right\}$$
$$+\lambda_1^{(n)}\|f^*\|_\psi^2. \tag{38}$$

Now, by the triangular inequality, the term $\|\hat{f}-f^*\|_\psi^2$ can be bounded as

$$\|\hat{f}-f^*\|_\psi^2 \le \left(\|\hat{f}\|_\psi+\|f^*\|_\psi\right)^2 \le 2\left(\|\hat{f}\|_\psi^2+\|f^*\|_\psi^2\right).$$

Thus, when $\lambda_1^{(n)} \ge \left(\frac{\alpha_2}{\alpha_1}\right)^2+\left(\frac{\beta_2}{\beta_1}\right)^2$, Eq. (38) yields

$$\frac{1}{2}\|\hat{f}-f^*\|_{L_2(\Pi)}^2 \le \frac{12\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2+\beta_1^2+\frac{M\log(M)}{n}\right)+2\lambda_1^{(n)}\|f^*\|_\psi^2.$$

Therefore by multiplying 2 to both sides, we have

$$\|\hat{f}-f^*\|_{L_2(\Pi)}^2 \le \frac{24\eta(t)^2\phi^2}{\kappa_M}\left(\alpha_1^2+\beta_1^2+\frac{M\log(M)}{n}\right)+4\lambda_1^{(n)}\|f^*\|_\psi^2.$$

This gives the assertion. □

Appendix D: Bounding the probabilities of $\mathscr{E}_1(t)$ and $\mathscr{E}_2(t')$

Here we derive bounds of the probabilities of the events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(t')$ (see Eq. (28) and Eq. (29) for their definitions). The goal of this section is to derive Lemmas 24 and 25.

Using Propositions 13 and 11, we obtain the following ratio type uniform bound.

Lemma 21. *Suppose that $\{\xi_i\}_{i=1}^n$ is a sequence of i.i.d. sub-Gaussian random variables satisfying $\mathrm{E}[e^{\xi_i t}] \le \exp(t^2/2)$ for all $t \in \mathbb{R}$. Under the Spectral Assumption (Assumption 3) and the Embedded Assumption (Assumption 5), there exists a constant C_{s_m} depending only on s_m, c and C_1 such that*

$$\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m:\|f_m\|_{\mathcal{H}_m}=1}\frac{|\frac{1}{n}\sum_{i=1}^n \xi_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)}\right] \le C_{s_m}.$$

Proof of Lemma 21. (i) First, we analyze the situation $0 < s_m < 1$. Let $\mathcal{H}_m(\delta) := \{f_m \in \mathcal{H}_m \mid \|f_m\|_{\mathcal{H}_m} = 1, \|f_m\|_{L_2(\Pi)} \le \delta\}$ and $z = 2^{1/s_m} > 1$. Define $\tau := s_m r_m$. Then by combining Propositions 10 and 11 with Assumption 5, we have

$$\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m:\|f_m\|_{\mathcal{H}_m}=1}\frac{|\frac{1}{n}\sum_{i=1}^n \xi_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)}\right]$$
$$\le \mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m(\tau)}\frac{|\frac{1}{n}\sum_{i=1}^n \xi_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)}\right]$$

$$
\begin{aligned}
&+\sum_{k=1}^{\infty}\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m(\tau z^k)\setminus\mathcal{H}_m(\tau z^{k-1})}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\right]\\
\leq& C'_{s_m}\frac{\frac{\tau^{1-s_m}\tilde{c}_{s_m}^{s_m}}{\sqrt{n}}}{3\frac{r_m^{-s_m}}{\sqrt{n}}s_m r_m}\vee\frac{\frac{C_1^{\frac{1-s_m}{1+s_m}}\tau^{\frac{(1-s_m)^2}{1+s_m}}\tilde{c}_{s_m}^{\frac{2s_m}{1+s_m}}}{n^{\frac{1}{1+s_m}}}}{3\frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}s_m r_m}\\
&+\sum_{k=1}^{\infty}C'_{s_m}\frac{\frac{z^{k(1-s_m)}\tau^{1-s_m}\tilde{c}_{s_m}^{s_m}}{\sqrt{n}}}{3\frac{r_m^{-s_m}}{\sqrt{n}}\tau z^{k-1}}\vee\frac{\frac{C_1^{\frac{1-s_m}{1+s_m}}z^{k\frac{(1-s_m)^2}{1+s_m}}\tau^{\frac{(1-s_m)^2}{1+s_m}}\tilde{c}_{s_m}^{\frac{2s_m}{1+s_m}}}{n^{\frac{1}{1+s_m}}}}{3\frac{r_m^{-\frac{s_m(3-s_m)}{1+s_m}}}{n^{\frac{1}{1+s_m}}}\tau z^{k-1}}\\
\leq&\frac{C'_{s_m}}{3}\left(s_m^{-s_m}\tilde{c}_{s_m}^{s_m}\vee s_m^{-3s_m}C_1^{\frac{1-s_m}{1+s_m}}\tilde{c}_{s_m}^{\frac{2s_m}{1+s_m}}\right)\left(1+\sum_{k=1}^{\infty}z^{1-ks_m}\vee z^{1-k\frac{s_m(3-s_m)}{1+s_m}}\right)\\
=&\frac{C'_{s_m}s_m^{-3s_m}}{3}\left(\tilde{c}_{s_m}^{s_m}\vee C_1^{\frac{1-s_m}{1+s_m}}\tilde{c}_{s_m}^{\frac{2s_m}{1+s_m}}\right)\left(1+\frac{z^{1-s_m}}{1-z^{-s_m}}\vee\frac{z^{1-\frac{s_m(3-s_m)}{1+s_m}}}{1-z^{-\frac{s_m(3-s_m)}{1+s_m}}}\right)\\
\leq&9C'_{s_m}\left(\tilde{c}_{s_m}^{s_m}\vee C_1^{\frac{1-s_m}{1+s_m}}\tilde{c}_{s_m}^{\frac{2s_m}{1+s_m}}\right)\left(1+\frac{z^{1-s_m}}{1-z^{-s_m}}\vee\frac{z^{1-\frac{s_m(3-s_m)}{1+s_m}}}{1-z^{-\frac{s_m(3-s_m)}{1+s_m}}}\right),
\end{aligned}
$$

where we used $s_m^{-s_m}\leq 3$ for $0<s_m$ in the last line. Thus by setting, $C_{s_m}=9C'_{s_m}\left(\tilde{c}_{s_m}^{s_m}\vee C_1^{\frac{1-s_m}{1+s_m}}\tilde{c}_{s_m}^{\frac{2s_m}{1+s_m}}\right)\left(1+\frac{z^{1-s_m}}{1-z^{-s_m}}\vee\frac{z^{1-\frac{s_m(3-s_m)}{1+s_m}}}{1-z^{-\frac{s_m(3-s_m)}{1+s_m}}}\right)$, we obtain the assertion.

(ii) Second, we analyze the situation $s_m=0$. In this situation, it is easy to see that, for any $f_m\in\mathcal{H}_m$, it holds that $f_m/U^{(m)}_{n,s_m}(f_m)\in\mathcal{H}_m$ and $\|f_m/U^{(m)}_{n,s_m}(f_m)\|_{L_2(\Pi)}\leq\sqrt{n}$. Therefore, by Propositions 12, we have that

$$
\begin{aligned}
&\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m:\|f_m\|_{\mathcal{H}_m}=1}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\right]\\
\leq&\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m:\|f_m\|_{L_2(\Pi)}\leq 1}\left|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)\right|\right]\leq\sqrt{\frac{c}{n}}\sqrt{n}=\sqrt{c}.
\end{aligned}
$$

Therefore, we obtain the assertion also for $s_m=0$. □

This lemma immediately gives the following corollary.

Corollary 22. *Suppose that $\{\xi_i\}_{i=1}^n$ is a sequence of i.i.d. sub-Gaussian random variables satisfying $\mathrm{E}[e^{\xi_i t}]\leq\exp(t^2/2)$ for all $t\in\mathbb{R}$. Under the Spectral Assumption (Assumption 3) and the Embedded Assumption (Assumption 5), there exists a constant C_{s_m} depending only on s_m, c and C_1 such that*

$$
\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\right]\leq C_{s_m}.
$$

Proof. By dividing the denominator and the numerator by the RKHS norm $\|f_m\|_{\mathcal{H}_m}$, we have

$$\begin{aligned}
&\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\right]\\
=&\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|/\|f_m\|_{\mathcal{H}_m}}{U^{(m)}_{n,s_m}(f_m)/\|f_m\|_{\mathcal{H}_m}}\right]\\
=&\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)/\|f_m\|_{\mathcal{H}_m}|}{U^{(m)}_{n,s_m}(f_m/\|f_m\|_{\mathcal{H}_m})}\right]\\
=&\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m:\|f_m\|_{\mathcal{H}_m}=1}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\right]\\
\le& C_{s_m}. \qquad (\because \text{Lemma 21})\square
\end{aligned}$$

Lemma 23. *Assume that $(\xi)_{i=1}^n$ is a sequence of Rademacher variables $(\sigma_i)_{i=1}^n$ or a sequence of i.i.d. standard normal variables. If $\frac{\log(M)}{\sqrt{n}}\le 1$, then under the Spectral Assumption (Assumption 3) and the Embedded Assumption (Assumption 5) there exists a constant $\tilde{C}_*$ depending only on $\{s_m\}_{m=1}^M$, c, C_1 such that*

$$\mathrm{E}\left[\max_m\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\xi_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\right]\le\tilde{C}_*.$$

Proof of Lemma 23. First, we assume that $(\xi)_{i=1}^n$ is a Rademacher sequence $(\sigma_i)_{i=1}^n$. Notice that the $L_2(\Pi)$-norm and the ∞-norm of $\frac{\sigma_i f_m(x_i)}{U^{(m)}_{n,s_m}(f_m)}$ can be evaluated by

$$\left\|\frac{\sigma_i f_m(x_i)}{U^{(m)}_{n,s_m}(f_m)}\right\|_{L_2(\Pi)}=\frac{\|f_m\|_{L_2(\Pi)}}{U^{(m)}_{n,s_m}(f_m)}\le\frac{\|f_m\|_{L_2(\Pi)}}{\sqrt{\frac{\log(M)}{n}}\|f_m\|_{L_2(\Pi)}}\le\sqrt{\frac{n}{\log(M)}}, \tag{39}$$

$$\left\|\frac{\sigma_i f_m(x_i)}{U^{(m)}_{n,s_m}(f_m)}\right\|_{\infty}=\frac{\|f_m\|_{\infty}}{U^{(m)}_{n,s_m}(f_m)}\le\frac{C_1\|f_m\|^{1-s_m}_{L_2(\Pi)}\|f_m\|^{s_m}_{\mathcal{H}_m}}{U^{(m)}_{n,s_m}(f_m)}\le\frac{C_1}{3}\sqrt{n}\le C_1\sqrt{n}, \tag{40}$$

where the second line is shown by using the relation (27). Let $C_*:=\max_m C_{s_m}$ where C_{s_m} is the constant appeared in Lemma 21. Thus Talagrand's inequality (Proposition 13) and Corollary 22 imply

$$\begin{aligned}
&P\left(\max_m\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\sigma_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\ge K\left[C_*+\sqrt{\frac{t}{\log(M)}}+\frac{C_1 t}{\sqrt{n}}\right]\right)\\
\le&\sum_{m=1}^M P\left(\sup_{f_m\in\mathcal{H}_m}\frac{|\frac{1}{n}\sum_{i=1}^n\sigma_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)}\ge K\left[C_*+\sqrt{\frac{t}{\log(M)}}+\frac{C_1 t}{\sqrt{n}}\right]\right)
\end{aligned}$$

$$\leq \sum_{m=1}^{M} P\left(\sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \geq K\left[C_{s_m} + \sqrt{\frac{t}{\log(M)}} + \frac{C_1 t}{\sqrt{n}} \right] \right)$$
$$\leq M e^{-t}.$$

By setting $t \leftarrow t + \log(M)$, we obtain

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \right.$$
$$\left. \geq K\left[C_* + \sqrt{\frac{t+\log(M)}{\log(M)}} + \frac{C_1(t+\log(M))}{\sqrt{n}} \right] \right) \leq e^{-t}$$

for all $t \geq 0$. Consequently the expectation of the max-sup term can be bounded as

$$\mathrm{E}\left[\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \right]$$
$$\leq K\left[C_* + 1 + \frac{C_1 \log(M)}{\sqrt{n}} \right]$$
$$+ \int_0^\infty K\left[C_* + \sqrt{\frac{t+1+\log(M)}{\log(M)}} + \frac{C_1(t+1+\log(M))}{\sqrt{n}} \right] e^{-t} \mathrm{d}t$$
$$\leq 2K\left[C_* + \sqrt{2} + \sqrt{\frac{\pi}{4\log(M)}} + \frac{C_1(2+\log(M))}{\sqrt{n}} \right] \leq \tilde{C}_*,$$

where we used $\sqrt{t+1+\log(M)} \leq \sqrt{t} + \sqrt{1+\log(M)}$ and $\int_0^\infty \sqrt{t} e^{-t} \mathrm{d}t = \sqrt{\frac{\pi}{4}}$, $\frac{\log(M)}{\sqrt{n}} \leq 1$, and $\tilde{C}_* = 2K[C_* + \sqrt{2} + \sqrt{\frac{\pi}{4}} + 3C_1]$.

The proof for the i.i.d. standard normal sequence $\{\xi_i\}_{i=1}^n$ is almost identical to that for the Rademacher sequence except that we use Lemma 18 instead of Talagrand's inequality (Proposition 13). □

Lemma 24. *Suppose the Basic Assumption (Assumption 1), the Spectral Assumption (Assumption 3) and the Embedded Assumption (Assumption 5) hold. Define* $\bar{\phi} = KL\left[2\tilde{C}_* + 1 + C_1\right]$. *If* $\frac{\log(M)}{\sqrt{n}} \leq 1$, *then the following holds*

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \geq \bar{\phi}\eta(t) \right) \leq e^{-t}.$$

Proof of Lemma 24. First, we assume a situation where $|\epsilon_i| \leq L$. By the contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12) and Lemma 23, we have

$$\mathrm{E}\left[\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \right] \leq 2\mathrm{E}\left[\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i \epsilon_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \right]$$

$$\leq 2L\tilde{C}_*,$$

where we used $\epsilon_i \leq L$ (Basic Assumption). Using this and Eq. (39) and Eq. (40), Talgrand's inequality (Proposition 13) gives

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \geq KL\left[2\tilde{C}_* + \sqrt{t} + \frac{C_1 t}{\sqrt{n}}\right]\right) \leq e^{-t}.$$

Thus we have

$$P\Bigg(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \geq KL\left[2\tilde{C}_* + 1 + C_1\right] \max\left(1, \sqrt{t}, \frac{t}{\sqrt{n}}\right)\Bigg) \leq e^{-t}.$$

Therefore by the definition of $\bar{\phi}$ and $\eta(t)$, we obtain the assertion.

Next, we consider the situation where ϵ_i is a Gaussian noise ($N(0, L^2)$). In this situation, we apply Lemma 18 instead of Talgrand's inequality (Proposition 13). Then, we obtain the assertion as in the bounded noise situation. □

Lemma 25. *Suppose the Basic Assumption (Assumption 1), the Spectral Assumption (Assumption 3) and the Embedded Assumption (Assumption 5) hold. Let* $\bar{\phi}' = K[2C_1\tilde{C}_* + C_1 + C_1^2]$. *Then, if* $\frac{\log(M)}{\sqrt{n}} \leq 1$, *we have for all* $t \geq 0$

$$\left|\left\|\sum_{m=1}^M f_m\right\|_n^2 - \left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2\right| \leq \phi'\sqrt{n}\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2 \eta(t),$$

for all $f_m \in \mathcal{H}_m$ $(m = 1, \ldots, M)$ *with probability* $1 - \exp(-t)$.

Proof of Lemma 25.

$$\begin{aligned}
&\mathrm{E}\left[\sup_{\substack{f_m \in \mathcal{H}_m \\ (m=1,\ldots,M)}} \frac{\left|\left\|\sum_{m=1}^M f_m\right\|_n^2 - \left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2\right|}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\right] \\
\leq& 2\mathrm{E}\left[\sup_{\substack{f_m \in \mathcal{H}_m \\ (m=1,\ldots,M)}} \frac{\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(\sum_{m=1}^M f_m(x_i))^2\right|}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\right] \\
\leq& \sup_{\substack{f_m \in \mathcal{H}_m \\ (m=1,\ldots,M)}} \frac{\left\|\sum_{m=1}^M f_m\right\|_\infty}{\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)} \times 2\mathrm{E}\left[\sup_{\substack{f_m \in \mathcal{H}_m \\ (m=1,\ldots,M)}} \frac{\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(\sum_{m=1}^M f_m(x_i))\right|}{\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)}\right],
\end{aligned} \tag{41}$$

where we used the contraction inequality in the last line (Ledoux and Talagrand, 1991, Theorem 4.12). Thus using Eq. (40), the RHS of the inequality (41) can be bounded as

$$
\begin{aligned}
&2C_1\sqrt{n}\mathrm{E}\left[\sup_{\substack{f_m\in\mathcal{H}_m\\(m=1,\ldots,M)}}\frac{\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(\sum_{m=1}^M f_m(x_i))\right|}{\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)}\right]\\
\leq&2C_1\sqrt{n}\mathrm{E}\left[\sup_{\substack{f_m\in\mathcal{H}_m\\(m=1,\ldots,M)}}\max_m\frac{\left|\frac{1}{n}\sum_{i=1}^n\sigma_i f_m(x_i)\right|}{U_{n,s_m}^{(m)}(f_m)}\right],
\end{aligned}
$$

where we used the relation

$$
\frac{\sum_m a_m}{\sum_m b_m}\leq\max_m\left(\frac{a_m}{b_m}\right) \tag{42}
$$

for all $a_m\geq 0$ and $b_m\geq 0$ with a convention $\frac{0}{0}=0$. By Lemma 23, the right hand side is upper bounded by $2C_1\sqrt{n}\tilde{C}_*$. Here we again apply Talagrand's concentration inequality, then we have

$$
\begin{aligned}
P\Bigg(&\sup_{\substack{f_m\in\mathcal{H}_m\\(m=1,\ldots,M)}}\frac{\left|\left\|\sum_{m=1}^M f_m\right\|_n^2-\left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2\right|}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\\
&\geq K\left[2C_1\tilde{C}_*\sqrt{n}+\sqrt{tn}C_1+C_1^2t\right]\Bigg)\leq e^{-t},
\end{aligned}
$$

where we substituted the following upper bounds of B and U.

$$
\begin{aligned}
B\leq&\sup_{\substack{f_m\in\mathcal{H}_m\\(m=1,\ldots,M)}}\mathrm{E}\left[\left(\frac{(\sum_{m=1}^M f_m)^2}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\right)^2\right]\\
\leq&\sup_{\substack{f_m\in\mathcal{H}_m\\(m=1,\ldots,M)}}\mathrm{E}\left[\frac{(\sum_{m=1}^M f_m)^2}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\frac{(\|\sum_{m=1}^M f_m\|_\infty)^2}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\right]\\
\overset{(40)}{\leq}&\sup_{\substack{f_m\in\mathcal{H}_m\\(m=1,\ldots,M)}}\frac{\left(\sum_{m=1}^M\|f_m\|_{L_2(\Pi)}\right)^2}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\frac{(\sum_{m=1}^M C_1\sqrt{n}U_{n,s_m}^{(m)}(f_m))^2}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2}\\
\overset{(39)}{\leq}&C_1^2n^2\frac{1}{\log(M)}\leq C_1^2n^2,
\end{aligned}
$$

where in the second inequality we used the relation

$$\mathrm{E}\left[\left(\sum_{m=1}^M f_m\right)^2\right] = \mathrm{E}\left[\sum_{m,m'=1}^M f_m f_{m'}\right] \leq \sum_{m,m'=1}^M \|f_m\|_{L_2(\Pi)}\|f_{m'}\|_{L_2(\Pi)} = \left(\sum_{m=1}^M \|f_m\|_{L_2(\Pi)}\right)^2$$

and in the third and forth inequality we used Eq. (40) and Eq. (39) with Eq.(42) respectively. Here we again use Eq. (39) with Eq.(42) to obtain

$$U = \sup_{\substack{f_m \in \mathcal{H}_m \\ (m=1,\ldots,M)}} \left\| \frac{(\sum_{m=1}^M f_m)^2}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2} \right\|_\infty \leq C_1^2 n.$$

Therefore the above inequality implies the following inequality

$$\sup_{\substack{f_m \in \mathcal{H}_m \\ (m=1,\ldots,M)}} \frac{\left| \|\sum_{m=1}^M f_m\|_n^2 - \|\sum_{m=1}^M f_m\|_{L_2(\Pi)}^2 \right|}{\left(\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m)\right)^2} \leq K\left[2C_1\tilde{C}_s + C_1 + C_1^2\right]\sqrt{n}\max(1,\sqrt{t},t/\sqrt{n}),$$

with probability $1-\exp(-t)$. Remind $\bar{\phi}' = K\left[2C_1\tilde{C}_* + C_1 + C_1^2\right]$, then we obtain the assertion. □

Appendix E: Proof of Theorem 5 (minimax learning rate)

Let *the δ-packing number* $Q(\delta, \mathcal{H}, L_2(\Pi))$ of a function class $\mathcal{H}$ be the largest number of functions $\{f_1, \ldots, f_Q\} \subseteq \mathcal{H}$ such that $\|f_i - f_j\|_{L_2(\Pi)} \geq \delta$ for all $i \neq j$.

Proof of Theorem 5. The proof utilizes the techniques developed by Raskutti et al. (2009, 2010) that applied the information theoretic technique developed by Yang and Barron (1999) to the MKL settings. To simplify the notation, we write $\mathcal{F} := \mathcal{H}_\psi(R)$, $N(\varepsilon, \mathcal{H}) := N(\varepsilon, \mathcal{H}, L_2(\Pi))$ and $Q(\varepsilon, \mathcal{H}) := Q(\varepsilon, \mathcal{H}, L_2(\Pi))$. It can be easily shown that $Q(2\varepsilon, \mathcal{F}) \leq N(2\varepsilon, \mathcal{F}) \leq Q(\varepsilon, \mathcal{F})$. Here due to Theorem 15 of Steinwart et al. (2009), Assumption 6 yields

$$\log N(\varepsilon, \tilde{\mathcal{H}}(1)) \sim \varepsilon^{-2s}. \tag{43}$$

We utilize the following inequality given by Lemma 3 of Raskutti et al. (2009):

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_\psi(R_p)} \mathrm{E}\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \geq \frac{\delta_n^2}{4}\left(1 - \frac{\log N(\varepsilon_n, \mathcal{F}) + n\varepsilon_n^2/2\sigma^2 + \log 2}{\log Q(\delta_n, \mathcal{F})}\right).$$

First we show the assertion for the ℓ_∞-norm ball: $\mathcal{H}_\psi(R) = \mathcal{H}_{\ell_\infty}(R) := \left\{ f = \sum_{m=1}^M f_m \;\middle|\; \max_{1\leq m\leq M} \|f_m\|_{\mathcal{H}_m} \leq R \right\}$. In this situation, there is a constant C that depends only s such that

$$\log Q(\delta, \mathcal{F}) \geq CM \log Q(\delta/\sqrt{M}, \widetilde{\mathcal{H}}(R)), \quad \log N(\varepsilon, \mathcal{F}) \leq M \log N(\varepsilon/\sqrt{M}, \widetilde{\mathcal{H}}(R)),$$

(this is shown in Lemma 5 of Raskutti et al. (2010), but we give the proof in Lemma 26 for completeness). Using this expression, the minimax-learning rate is bounded as

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_{\ell_p}(R_p)} \mathrm{E}\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \geq \frac{\delta_n^2}{4}\left(1 - \frac{M \log N(\varepsilon_n/\sqrt{M}, \widetilde{\mathcal{H}}(R)) + n\varepsilon_n^2/2\sigma^2 + \log 2}{CM \log Q(\delta_n/\sqrt{M}, \widetilde{\mathcal{H}}(R))}\right).$$

Here we choose ε_n and δ_n to satisfy the following relations:

$$\frac{n}{2\sigma^2}\varepsilon_n^2 \leq M \log N\left(\varepsilon_n/\sqrt{M}, \widetilde{\mathcal{H}}(R)\right), \tag{44}$$

$$M \log N\left(\varepsilon_n/\sqrt{M}, \widetilde{\mathcal{H}}(R)\right) \geq \log 2, \tag{45}$$

$$4 \log N\left(\epsilon_n/\sqrt{M}, \widetilde{\mathcal{H}}(R)\right) \leq C \log Q\left(\delta_n/\sqrt{M}, \widetilde{\mathcal{H}}(R)\right). \tag{46}$$

With ε_n and δ_n that satisfy the above relations (44) and (46), we have

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_{\ell_p}(R_p)} \mathrm{E}\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \geq \frac{\delta_n^2}{16}. \tag{47}$$

By Eq. (43), the relation (44) can be rewritten as

$$\frac{n}{2\sigma^2}\varepsilon_n^2 \leq CM\left(\frac{\varepsilon_n}{R\sqrt{M}}\right)^{-2s}.$$

It is sufficient to impose

$$\varepsilon_n^2 \leq Cn^{-\frac{1}{1+s}} M R^{\frac{2s}{1+s}}, \tag{48}$$

with a constant C. Since we have assumed that $n > \frac{\bar{c}^2 M^2}{R^2 \|\mathbf{1}\|_{\psi^*}^2}$ $(= \frac{1}{R^2}$ for $\|\cdot\|_\psi = \|\cdot\|_{\ell_\infty})$, the conditions (45) can be satisfied if the constant C in Eq. (48) is taken sufficiently small so that we have

$$\log 2 \leq \log N(\varepsilon_n/\sqrt{M}, \widetilde{\mathcal{H}}(R)) \sim \left(\frac{\varepsilon_n}{R\sqrt{M}}\right)^{-2s}. \tag{49}$$

The relation (46) can be satisfied by taking $\delta_n = c\varepsilon_n$ with an appropriately chosen constant c. Thus Eq. (47) gives

$$\min_{\hat{f}} \max_{f^* \in \mathcal{H}_{\ell_p}(R_p)} \mathrm{E}\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \geq Cn^{-\frac{1}{1+s}} M R^{\frac{2s}{1+s}}, \tag{50}$$

with a constant C. This gives the assertion for $p = \infty$.

Finally we show the assertion for general isotropic ψ-norm $\|\cdot\|_\psi$. To show that, we prove that $\mathcal{H}_{\ell_\infty}(R\|\mathbf{1}\|_{\psi^*}/(\bar{c}M)) \subset \mathcal{H}_\psi(R)$. This is true if $\frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}\mathbf{1} \in \mathcal{H}_\psi(R)$

because of the second condition of the definition (15) of isotropic property. By the isotropic property, the ψ-norm of $\frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}\mathbf{1}$ is bounded as

$$\left\|\frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}\mathbf{1}\right\|_{\psi} = \frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}\|\mathbf{1}\|_{\psi} \overset{\text{isotropic}}{\leq} \frac{R}{\bar{c}M}\bar{c}M = R.$$

Thus we have $\frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}\mathbf{1} \in \mathcal{H}_\psi(R)$ and thus $\mathcal{H}_{\ell_\infty}(R\|\mathbf{1}\|_{\psi^*}/(\bar{c}M)) \subset \mathcal{H}_\psi(R)$. Therefore we have

$$\begin{aligned}\min_{\hat{f}} \max_{f^* \in \mathcal{H}_\psi(R)} \mathrm{E}\|\hat{f} - f^*\|^2_{L_2(\Pi)} &\geq \min_{\hat{f}} \max_{f^* \in \mathcal{H}_{\ell_\infty}(R\|\mathbf{1}\|_{\psi^*}/(\bar{c}M))} \mathrm{E}\|\hat{f} - f^*\|^2_{L_2(\Pi)} \\ &\geq Cn^{-\frac{1}{1+s}} M \left(\frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}\right)^{\frac{2s}{1+s}}, \qquad (\because \text{Eq. (50)}).\end{aligned}$$

Note that due to the condition $n > \frac{\bar{c}^2 M^2}{R^2\|\mathbf{1}\|^2_{\psi^*}}$, Eq. (50) is still valid under the condition that $\frac{R\|\mathbf{1}\|_{\psi^*}}{\bar{c}M}$ is substituted into R in Eq. (50) (more precisely, Eq. (49) is valid). Resetting $C \leftarrow C\bar{c}^{-\frac{2s}{1+s}}$, we obtain the assertion. □

Lemma 26. *There is a constant C such that*

$$\log Q(\delta, \mathcal{H}_{\ell_\infty}(R)) \geq CM \log Q(\delta/\sqrt{M}, \widetilde{\mathcal{H}}(R)),$$

for sufficiently small δ.

Proof. The proof is analogous to that of Lemma 5 in Raskutti et al. (2010). We describe the outline of the proof. Let $N = Q(\sqrt{2}\delta/\sqrt{M}, \widetilde{\mathcal{H}}(R))$ and $\{f_m^1, \ldots, f_m^N\}$ be a $\sqrt{2}\delta/\sqrt{M}$-packing of $\mathcal{H}_m(R)$. Then we can construct a function class Υ as

$$\Upsilon = \left\{ f^{\boldsymbol{j}} = \sum_{m=1}^M f_m^{j_m} \mid \boldsymbol{j} = (j_1, \ldots, j_M) \in \{1, \ldots, N\}^M \right\}.$$

We denote by $[N] := \{1, \ldots, N\}$. For two functions $f^{\boldsymbol{j}}, f^{\boldsymbol{j}'} \in \Upsilon$, we have by the construction

$$\|f^{\boldsymbol{j}} - f^{\boldsymbol{j}'}\|^2_{L_2(\Pi)} = \sum_{m=1}^M \|f_m^{j_m} - f_m^{j'_m}\|^2_{L_2(\Pi)} \geq \frac{2\delta^2}{M} \sum_{m=1}^M \mathbf{1}[j_m \neq j'_m].$$

Thus, it suffices to construct a sufficiently large subset $A \subset [N]^M$ such that all different pairs $\boldsymbol{j}, \boldsymbol{j}' \in A$ have at least $M/2$ of Hamming distance $d_H(\boldsymbol{j}, \boldsymbol{j}') := \sum_{m=1}^M \mathbf{1}[j_m \neq j'_m]$.

Now we define $d_H(A, \boldsymbol{j}) := \min_{\boldsymbol{j}' \in A} d_H(\boldsymbol{j}', \boldsymbol{j})$. If $|A|$ satisfies

$$\left| \left\{ \boldsymbol{j} \in [N]^M \;\middle|\; d_H(A, \boldsymbol{j}) \leq \frac{M}{2} \right\} \right| < |[N]^M| = N^M, \tag{51}$$

then there exists a member $\boldsymbol{j}' \in [N]^M$ such that $\boldsymbol{j}'$ is more than $\frac{M}{2}$ away from A with respect to d_H, i.e. $d_H(A, \boldsymbol{j}') > \frac{M}{2}$. That is, we can add $\boldsymbol{j}'$ to A as long as Eq. (51) holds. Now since

$$\left| \left\{ \boldsymbol{j} \in [N]^M \;\middle|\; d_H(A, \boldsymbol{j}) \leq \frac{M}{2} \right\} \right| \leq |A| \binom{M}{M/2} N^{M/2}, \tag{52}$$

Eq. (51) holds as long as A satisfies

$$|A| \leq \frac{1}{2} \frac{N^M}{\binom{M}{M/2} N^{M/2}} =: Q^*.$$

The logarithm of Q^* can be evaluated as follows

$$\begin{aligned} \log Q^* &= \log \left(\frac{1}{2} \frac{N^M}{\binom{M}{M/2} N^{M/2}} \right) = M \log N - \log 2 - \log \binom{M}{M/2} - \frac{M}{2} \log N \\ &\geq \frac{M}{2} \log N - \log 2 - \log 2^M \geq \frac{M}{2} \log \frac{N}{16}. \end{aligned}$$

There exists a constant C such that $N = Q(\sqrt{2}\delta/\sqrt{M}, \widetilde{\mathcal{H}}(R)) \geq CQ(\delta/\sqrt{M}, \widetilde{\mathcal{H}}(R))$ because $\log Q(\delta, \widetilde{\mathcal{H}}(R)) \sim \left(\frac{\delta}{R}\right)^{-2s}$. Thus we obtain the assertion for sufficiently large N. □

Appendix F: Proof of technical lemmas

F.1. Proof of Lemma 2

Remind that Eq. (7) gives

$$\begin{aligned} &\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\ &= \mathcal{O}_p \left(\min_{\substack{\{r_m\}_{m=1}^M : \\ r_m > 0}} \left\{ \alpha_1^2 + \beta_1^2 + \left[\left(\frac{\alpha_2}{\alpha_1} \right)^2 + \left(\frac{\beta_2}{\beta_1} \right)^2 \right] \|f^*\|_\psi^2 + \frac{M \log(M)}{n} \right\} \right). \end{aligned} \tag{53}$$

We derive an upper bound of the right hand side by adding a constraint $r_m = r$ $(\forall m)$. Since $s_m = s$ $(\forall m)$, under the constraint $r_m = r$ $(\forall m)$ we have

$$\begin{aligned} \frac{\alpha_2}{\alpha_1} &= \frac{3 \frac{s r^{1-s}}{\sqrt{n}} \|\mathbf{1}\|_{\psi^*}}{3\sqrt{M \frac{r^{-2s}}{n}}} = \frac{1}{\sqrt{M}} sr \|\mathbf{1}\|_{\psi^*}, \\ \frac{\beta_2}{\beta_1} &= \frac{3 \frac{s r^{\frac{(1-s)^2}{1+s}}}{n^{\frac{1}{1+s}}} \|\mathbf{1}\|_{\psi^*}}{3\sqrt{M \frac{r^{-\frac{2s(3-s)}{1+s}}}{n^{\frac{2}{1+s}}}}} = \frac{1}{\sqrt{M}} sr \|\mathbf{1}\|_{\psi^*}, \end{aligned}$$

Thus $\frac{\alpha_2}{\alpha_1} = \frac{\beta_2}{\beta_1}$, and Eq. (53) becomes

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(\min_{\substack{r>0,\\ r_m = r}} \left\{ \alpha_1^2 + \beta_1^2 + 2\frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2 + \frac{M\log(M)}{n} \right\} \right). \tag{54}$$

By the definition, we see that the first two terms are monotonically decreasing function with respect to r and the third term is monotonically increasing function. The minimum of the right hand side is attained by balancing $\alpha_1^2 + \beta_1^2$ and $2\frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2$. Since $\alpha_1^2 + \beta_1^2 \leq 2\max\left(\alpha_1^2, \beta_1^2\right)$, Eq. (54) indicates that

$$\begin{aligned} &\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\ &\quad \leq \mathcal{O}_p\left(\min_{\substack{r>0,\\ r_m = r}} \left\{ 2\max\left(\alpha_1^2, \beta_1^2\right) + 2\frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2 + \frac{M\log(M)}{n} \right\} \right). \end{aligned} \tag{55}$$

To balance the first term and the second term, we need to consider two situations: $\alpha_1^2 = \frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2$ or $\beta_1^2 = \frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2$.

First we balance the terms α_1^2 and $\frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2$ under the restriction that $r_m = r$ $(\forall m)$:

$$\begin{aligned} & \alpha_1^2 = \frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2 \\ \Leftrightarrow \quad & 9M\frac{r^{-2s}}{n} = \frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2 \\ \Leftrightarrow \quad & r^{-1} = (s/3)^{\frac{1}{1+s}} M^{-\frac{1}{1+s}} n^{\frac{1}{2(1+s)}} \left(\|\mathbf{1}\|_{\psi^*} \|f^*\|_{\psi}\right)^{\frac{1}{1+s}}. \end{aligned} \tag{56}$$

For this r, we obtain

$$\begin{aligned} \alpha_1^2 &= 9M\frac{r^{-2s}}{n} \\ &= 9^{\frac{1}{1+s}} s^{\frac{2s}{1+s}} M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} \left(\|\mathbf{1}\|_{\psi^*} \|f^*\|_{\psi}\right)^{\frac{2s}{1+s}} \\ &\leq 9M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} \left(\|\mathbf{1}\|_{\psi^*} \|f^*\|_{\psi}\right)^{\frac{2s}{1+s}}, \end{aligned} \tag{57}$$

where we used $s^{\frac{2s}{1+s}} \leq 1$ and $9^{\frac{1}{1+s}} \leq 9$ in the last inequality.

Next we balance the terms β_1^2 and $\frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2$ under the restriction that $r_m = r$ $(\forall m)$:

$$\begin{aligned} & \beta_1^2 = \frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2 \\ \Leftrightarrow \quad & 9M\frac{r^{-\frac{2s(3-s)}{1+s}}}{n^{\frac{2}{1+s}}} = \frac{1}{M} s^2 r^2 \|\mathbf{1}\|_{\psi^*}^2 \|f^*\|_{\psi}^2 \\ \Leftrightarrow \quad & r^{-1} = (s/3)^{\frac{1+s}{1+4s-s^2}} M^{-\frac{1+s}{1+4s-s^2}} n^{\frac{1}{1+4s-s^2}} \left(\|\mathbf{1}\|_{\psi^*} \|f^*\|_{\psi}\right)^{\frac{1+s}{1+4s-s^2}}. \end{aligned}$$

For this r, we obtain

$$\begin{aligned}\beta_1^2 &= 9M\frac{r^{-\frac{2s(3-s)}{1+s}}}{n^{\frac{2}{1+s}}}\\ &= 9^{\frac{1+s}{1+4s-s^2}} s^{\frac{2s(3-s)}{1+4s-s^2}} M^{-\frac{1-2s+s^2}{1+4s-s^2}} n^{-\frac{2}{1+4s-s^2}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s(3-s)}{1+4s-s^2}}\\ &\le 9M^{\frac{1-2s+s^2}{1+4s-s^2}} n^{-\frac{2}{1+4s-s^2}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s(3-s)}{1+4s-s^2}},\end{aligned}$$

where we used $s^{\frac{2s(3-s)}{1+4s-s^2}} \le 1$ and $9^{\frac{1+s}{1+4s-s^2}} \le 9$ in the last inequality.

Therefore the right hand side of Eq. (55) is further bounded as

$$\begin{aligned}&\|\hat{f} - f^*\|_{L_2(\Pi)}^2\\ &\le \mathcal{O}_p\Bigg(4\max\Bigg\{9M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}},\\ &\qquad 9M^{\frac{1-2s+s^2}{1+4s-s^2}} n^{-\frac{2}{1+4s-s^2}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s(3-s)}{1+4s-s^2}}\Bigg\} + \frac{M\log(M)}{n}\Bigg)\\ &= \mathcal{O}_p\Bigg(M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} +\\ &\qquad M^{\frac{(1-s)^2}{1+4s-s^2}} n^{-\frac{2}{1+4s-s^2}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s(3-s)}{1+4s-s^2}} + \frac{M\log(M)}{n}\Bigg).\end{aligned}$$

Finally, if $n \ge (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi/M)^{\frac{4s}{1-s}}$, the first term of the right hand side of this bound is not less than the second term:

$$M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} \ge M^{\frac{(1-s)^2}{1+4s-s^2}} n^{-\frac{2}{1+4s-s^2}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s(3-s)}{1+4s-s^2}}.$$

More precisely, with r given in Eq. (56), the upper bound (57) of α_1 gives that, for $n \ge (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi/M)^{\frac{4s}{1-s}}$, we have

$$\begin{aligned}&\sqrt{n}\max\left\{\alpha_1^2, \beta_1^2, \frac{M\log(M)}{n}\right\}\\ &\le \sqrt{n}9M^{1-\frac{2s}{1+s}} n^{-\frac{1}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} \vee \frac{M\log(M)}{\sqrt{n}}\\ &= 9\left(\frac{M}{\sqrt{n}}\right)^{\frac{1-s}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} \vee \frac{M\log(M)}{\sqrt{n}}.\end{aligned}$$

Thus by setting $\lambda_1^{(n)} = 18M^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} \|\mathbf{1}\|_{\psi^*}^{\frac{2s}{1+s}} \|f^*\|_\psi^{-\frac{2}{1+s}} \ge \left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2$, then Theorem 1 gives that for all n and t' that satisfy $\frac{\log(M)}{\sqrt{n}} \le 1$ and $\frac{4\phi}{\kappa_M}\left\{9\left(\frac{M}{\sqrt{n}}\right)^{\frac{1-s}{1+s}} (\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} \vee \frac{M\log(M)}{\sqrt{n}}\right\}\eta(t') \le \frac{1}{12}$ and for all $t \ge 1$, we

have

$$\begin{aligned}\|\hat{f}-f^*\|_{L_2(\Pi)}^2 \leq & \frac{24\eta(t)^2\phi^2}{\kappa_M}\left(18M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}}+\frac{M\log(M)}{n}\right)\\ &+4\times 18M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}} \qquad (58)\\ \leq & C\eta(t)^2\left(M^{1-\frac{2s}{1+s}}n^{-\frac{1}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}}+\frac{M\log(M)}{n}\right),\end{aligned}$$

with probability $1-\exp(-t)-\exp(-t')$ where C is a sufficiently large constant depending on ϕ and κ_M. Finally notice that the condition $\frac{4\phi}{\kappa_M}\left\{9\left(\frac{M}{\sqrt{n}}\right)^{\frac{1-s}{1+s}}(\|\mathbf{1}\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}}\vee\frac{M\log(M)}{\sqrt{n}}\right\}\eta(t')\leq\frac{1}{12}$ automatically gives $\frac{\log(M)}{\sqrt{n}}\leq 1$, thus we can drop the condition $\frac{\log(M)}{\sqrt{n}}\leq 1$. Then we obtain the assertion.

F.2. Proof of Lemma 3

We assume $1<p<\infty$ and $1<q<\infty$. The proof for the situations $p=1,\infty$ or $q=1,\infty$ is straight forward. First applying Hölder's inequality twice, we obtain

$$\begin{aligned}\langle \boldsymbol{b},\boldsymbol{a}\rangle &= \sum_{j=1}^{M'}\sum_{k=1}^{M_j} b_{j,k}a_{j,k}\\ &\leq \sum_{j=1}^{M'}\left\{\left(\sum_{k=1}^{M_j}|b_{j,k}|^{p^*}\right)^{\frac{1}{p^*}}\left(\sum_{k=1}^{M_j}|a_{j,k}|^{p}\right)^{\frac{1}{p}}\right\} \quad (\because \text{Hölder's inequality})\\ &\leq \left\{\sum_{j=1}^{M'}\left(\sum_{k=1}^{M_j}|b_{j,k}|^{p^*}\right)^{\frac{q^*}{p^*}}\right\}^{\frac{1}{q^*}}\left\{\sum_{j=1}^{M'}\left(\sum_{k=1}^{M_j}|a_{j,k}|^{p}\right)^{\frac{q}{p}}\right\}^{\frac{1}{q}}\\ &\quad (\because \text{Hölder's inequality}).\end{aligned}$$

Therefore we obtain that

$$\|\boldsymbol{b}\|_{\psi^*}\leq\left\{\sum_{j=1}^{M'}(\sum_{k=1}^{M_j}|b_{j,k}|^{p^*})^{\frac{q^*}{p^*}}\right\}^{\frac{1}{q^*}}. \qquad (59)$$

On the other hand, if we set

$$a_{j,k}=b_{j,k}^{\frac{1}{p-1}}\frac{(\sum_{k=1}^{M_j}b_{j,k}^{p^*})^{\frac{q^*}{p^*}-1}}{\{\sum_{j'=1}^{M'}(\sum_{k=1}^{M_{j'}}b_{j',k}^{p^*})^{\frac{q^*}{p^*}}\}^{\frac{1}{q}}},$$

then we have

$$\|\boldsymbol{a}\|_{\psi} = \left\{ \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} b_{j,k}^{\frac{p}{p-1}} \right)^{\frac{q}{p}} \left(\sum_{k=1}^{M_j} b_{j,k}^{p^*} \right)^{q(\frac{q^*}{p^*}-1)} \right\}^{\frac{1}{q}} \frac{1}{\{\sum_{j'=1}^{M'} (\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*})^{\frac{q^*}{p^*}}\}^{\frac{1}{q}}}$$

$$= \left\{ \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} b_{j,k}^{\frac{p}{p-1}} \right)^{q\left(\frac{1}{p}-1+\frac{q^*}{p^*}\right)} \right\}^{\frac{1}{q}} \frac{1}{\{\sum_{j'=1}^{M'} (\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*})^{\frac{q^*}{p^*}}\}^{\frac{1}{q}}}$$

$$= \left\{ \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} b_{j,k}^{\frac{p}{p-1}} \right)^{\frac{q^*}{q^*-1}\left(\frac{q^*-1}{p^*}\right)} \right\}^{\frac{1}{q}} \frac{1}{\{\sum_{j'=1}^{M'} (\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*})^{\frac{q^*}{p^*}}\}^{\frac{1}{q}}} = 1,$$

and

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{j=1}^{M'} \left\{ \left(\sum_{k=1}^{M_j} b_{j,k}^{1+\frac{1}{p-1}} \right) \left(\sum_{k=1}^{M_j} b_{j,k}^{p^*} \right)^{\frac{q^*}{p^*}-1} \right\} \frac{1}{\{\sum_{j'=1}^{M'} (\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*})^{\frac{q^*}{p^*}}\}^{\frac{1}{q}}}$$

$$= \sum_{j=1}^{M'} \left(\sum_{k=1}^{M_j} b_{j,k}^{p^*} \right)^{\frac{q^*}{p^*}} \frac{1}{\{\sum_{j'=1}^{M'} (\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*})^{\frac{q^*}{p^*}}\}^{\frac{1}{q}}}$$

$$= \left\{ \sum_{j'=1}^{M'} \left(\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*} \right)^{\frac{q^*}{p^*}} \right\}^{\frac{1}{q^*}}.$$

Therefore we obtain

$$\|\boldsymbol{b}\|_{\psi^*} \geq \left\{ \sum_{j'=1}^{M'} \left(\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*} \right)^{\frac{q^*}{p^*}} \right\}^{\frac{1}{q^*}}. \tag{60}$$

Combining Eqs.(60),(60), we have $\|\boldsymbol{b}\|_{\psi^*} = \left\{ \sum_{j'=1}^{M'} \left(\sum_{k=1}^{M_{j'}} b_{j',k}^{p^*} \right)^{\frac{q^*}{p^*}} \right\}^{\frac{1}{q^*}}$. Thus we obtain the assertion.

F.3. Proof of Lemma 6

Suppose that $\|\cdot\|_{\psi}$ is the ℓ_p-norm $\|\cdot\|_{\ell_p}$, and remind that $\mathbf{1}_d = (\underbrace{1, \ldots, 1}_{d \text{ elements}}, 0, \ldots, 0)^\top$.

Since we can evaluate

$$\alpha_1 = 3\left(\frac{dr_1^{-2s} + M - d}{n}\right)^{\frac{1}{2}}, \alpha_2 = 3\frac{sr_1^{1-s}}{\sqrt{n}}\|\mathbf{1}_d\|_{\psi^*},$$
$$\beta_1 = 3\left(\frac{dr_1^{-\frac{2s(3-s)}{1+s}} + M - d}{n^{\frac{2}{1+s}}}\right)^{\frac{1}{2}}, \beta_2 = 3\frac{sr_1^{\frac{(1-s)^2}{1+s}}}{n^{\frac{1}{1+s}}}\|\mathbf{1}_d\|_{\psi^*},$$

then we have

$$\left(\frac{\alpha_2}{\alpha_1}\right)^2 = \frac{\frac{s^2 r_1^{2(1-s)}}{n}\|\mathbf{1}_d\|_{\psi^*}^2}{\frac{dr_1^{-2s}+M-d}{n}} \simeq \min\left\{\frac{s^2 r_1^2}{d}, \frac{s^2 r_1^{2(1-s)}}{M-d}\right\}\|\mathbf{1}_d\|_{\psi^*}^2,$$

and

$$\left(\frac{\beta_2}{\beta_1}\right)^2 = \frac{\frac{s^2 r_1^{\frac{2(1-s)^2}{1+s}}}{n^{\frac{2}{1+s}}}\|\mathbf{1}_d\|_{\psi^*}^2}{\frac{dr_1^{-\frac{2s(3-s)}{1+s}}+M-d}{n^{\frac{2}{1+s}}}} \simeq \min\left\{\frac{s^2 r_1^2}{d}, \frac{s^2 r_1^{\frac{2(1-s)^2}{1+s}}}{M-d}\right\}\|\mathbf{1}_d\|_{\psi^*}^2.$$

Suppose $dr_1^{-2s} \geq M-d$ and $dr_1^{-\frac{2s(3-s)}{1+s}} \geq M-d$, then we have $\alpha_1^2 \simeq dr_1^{-2s}n^{-1}$, $\beta_1^2 \simeq dr_1^{-\frac{2s(3-s)}{1+s}} n^{-\frac{2}{1+s}}$, $\left(\frac{\alpha_2}{\alpha_1}\right)^2 \simeq \frac{s^2 r_1^2}{d}\|\mathbf{1}_d\|_{\psi^*}^2$ and $\left(\frac{\beta_2}{\beta_1}\right)^2 \simeq \frac{s^2 r_1^2}{d}\|\mathbf{1}_d\|_{\psi^*}^2$. Thus the minimization problem in Eq. (7) with the constraint for r_1 becomes

$$\min_{\substack{r_1>0:\\ dr_1^{-2s}\geq M-d,\\ dr_1^{-\frac{2s(3-s)}{1+s}}\geq M-d}} \left\{\alpha_1^2 + \beta_1^2 + \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2\right\}$$
$$\simeq \min_{\substack{r_1>0:\\ dr_1^{-2s}\geq M-d,\\ dr_1^{-\frac{2s(3-s)}{1+s}}\geq M-d}} \left\{dr_1^{-2s}n^{-1} + dr_1^{-\frac{2s(3-s)}{1+s}} n^{-\frac{2}{1+s}} + \frac{r_1^2}{d}\|\mathbf{1}_d\|_{\psi^*}^2\|f^*\|_\psi^2\right\}. \quad (61)$$

If we neglect the constraints $dr_1^{-2s} \geq M-d$ and $dr_1^{-\frac{2s(3-s)}{1+s}} \geq M-d$, the minimum is attained at r_1 (up to a constant factor) that satisfies $\max\{dr_1^{-2s}n^{-1}, dr_1^{-\frac{2s(3-s)}{1+s}} n^{-\frac{2}{1+s}}\} = \frac{1}{d}r_1^2\|\mathbf{1}_d\|_{\psi^*}^2\|f^*\|_\psi^2$, i.e.

$$r_1 = \max\left\{n^{-\frac{1}{2(1+s)}}\left(\frac{\|\mathbf{1}_d\|_{\psi^*}\|f^*\|_\psi}{d}\right)^{-\frac{1}{1+s}}, n^{-\frac{1}{1+4s-s^2}}\left(\frac{\|\mathbf{1}_d\|_{\psi^*}\|f^*\|_\psi}{d}\right)^{-\frac{1+s}{1+4s-s^2}}\right\}.$$

Therefore if $n \geq \left(\frac{\|\mathbf{1}_d\|_{\psi^*}\|f^*\|_\psi}{d}\right)^{\frac{4s}{1-s}}$ (this is satisfied because $M = \|f^*\|_{\ell_1} \geq \|f^*\|_{\ell_p} \geq \|f^*\|_{\ell_\infty} = 1$, $\|\mathbf{1}_d\|_{\psi^*} \leq \|\mathbf{1}_d\|_1 \leq d$ and $n \geq M^{\frac{4s}{1-s}}$ is imposed),

then the minimum is attained at $r_1 = n^{-\frac{1}{2(1+s)}} \left(\frac{\|\mathbf{1}_d\|_{\psi^*}\|f^*\|_\psi}{d}\right)^{-\frac{1}{1+s}}$. Finally the condition $n \geq (M\log(M))^{\frac{1+s}{s}}$ yields that $dr_1^{-2s} \geq M-d$ and $dr_1^{-\frac{2s(3-s)}{1+s}} \geq M-d$ for $r_1 = n^{-\frac{1}{2(1+s)}} \left(\frac{\|\mathbf{1}_d\|_{\psi^*}\|f^*\|_\psi}{d}\right)^{-\frac{1}{1+s}}$. Therefore the constraints for r_1 in Eq. (61) can be removed. Summarizing the above discussions, we obtain

$$\min_{\substack{\{r_m\}_{m=1}^M: \\ r_m>0}} \left\{\alpha_1^2 + \beta_1^2 + \left[\left(\frac{\alpha_2}{\alpha_1}\right)^2 + \left(\frac{\beta_2}{\beta_1}\right)^2\right]\|f^*\|_\psi^2\right\} \simeq n^{-\frac{1}{1+s}} \frac{(\|\mathbf{1}_d\|_{\psi^*}\|f^*\|_\psi)^{\frac{2s}{1+s}}}{d^{\frac{s-1}{1+s}}}.$$

Thus we obtain the following bound:

$$\|\hat{f}^{(p)} - f^*\|_{L_2(\Pi)}^2 \leq \mathcal{O}_p\left(n^{-\frac{1}{1+s}} \frac{(d^{1-\frac{1}{p}}\|f^*\|_\psi)^{\frac{2s}{1+s}}}{d^{\frac{s-1}{1+s}}} + \frac{M\log(M)}{n}\right).$$

Now since $n \geq (M\log(M))^{\frac{1+s}{s}}$, the above convergence rates can be simplified as

$$\|\hat{f}^{(p)} - f^*\|_{L_2(\Pi)}^2 = \mathcal{O}_p\left(n^{-\frac{1}{1+s}} d^{1-\frac{2s}{p(1+s)}} \|f^*\|_\psi^{\frac{2s}{1+s}}\right).$$

In particular, if $d = M^{b_1}$ and $\|f_m^*\|_{\mathcal{H}_m} = m^{-b_2}$, then it holds that

$$\begin{aligned}
\|\hat{f}^{(1)} - f^*\|_{L_2(\Pi)}^2 &= \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{\frac{2s}{1+s}(1-b_2)+\frac{1-s}{1+s}b_1}\right), \\
\|\hat{f}^{(\infty)} - f^*\|_{L_2(\Pi)}^2 &= \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{b_1}\right), \\
\|\hat{f}^{(b_2^{-1})} - f^*\|_{L_2(\Pi)}^2 &= \mathcal{O}_p\left(n^{-\frac{1}{1+s}} M^{b_1(1-\frac{2sb_2}{1+s})} \log(M)^{b_2\frac{2s}{1+s}}\right).
\end{aligned}$$

This gives the assertion.

F.4. Proof of Lemma 7 (derivation of local Rademacher complexity)

For $f \in \mathcal{H}^{\oplus M}$, we define

$$U_{n,*}(f) := \alpha_1 \frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}} + \alpha_2\|f\|_\psi + \beta_1 \frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}} + \beta_2\|f\|_\psi + \sqrt{\frac{M\log(M)}{n}} \frac{\|f\|_{L_2(\Pi)}}{\sqrt{\kappa_M}}.$$

Then by Eq. (35) we obtain

$$\sum_{m=1}^M U_{n,s_m}^{(m)}(f_m) \leq U_{n,*}(f).$$

We know that there exists a constant $\tilde{\phi}$ such that

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{U_{n,s_m}^{(m)}(f_m)} \geq \tilde{\phi}\eta(t)\right) \leq e^{-t}, \tag{62}$$

(see Lemma 24). Let $\bar{\eta}(t) := \max\{\sqrt{t}, t/n\}$, and the event $\mathcal{S}_t$ be

$$\mathcal{S}_t := \left\{ \tilde{\phi}\bar{\eta}(t) \leq \max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{U^{(m)}_{n,s_m}(f_m)} \leq \tilde{\phi}\bar{\eta}(t+1) \right\}.$$

Then, by Eq. (62), we have $P(\mathcal{S}_t) \leq e^{-t}$ for $t \geq 1$. Using this relation, we obtain the following upper bound of the local Rademacher complexity:

$$\begin{aligned}
&R_n(\mathcal{H}^{(r)}_\psi(R)) \\
=&\mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)\right] \\
=&\sum_{t=0}^\infty \mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \mid \mathcal{S}_t\right] P(\mathcal{S}_t) \\
\leq&\mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \mid \mathcal{S}_0\right] + \sum_{t=1}^\infty \mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i) \mid \mathcal{S}_t\right] P(\mathcal{S}_t) \\
\leq&\mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \sum_{m=1}^M \tilde{\phi} U^{(m)}_{n,s_m}(f_m) \mid S_0\right] \\
&+ \sum_{t=1}^\infty \mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \sum_{m=1}^M \tilde{\phi}\eta(t+1) U^{(m)}_{n,s_m}(f_m) \mid \mathcal{S}_t\right] e^{-t} \\
\leq&\mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \tilde{\phi} U_{n,*}(f) \mid S_0\right] + \sum_{t=1}^\infty \mathrm{E}\left[\sup_{f\in\mathcal{H}^{(r)}_\psi(R)} \tilde{\phi}\eta(t+1) U_{n,*}(f) \mid \mathcal{S}_t\right] e^{-t} \\
\leq&\tilde{\phi}\left(\alpha_1 \frac{r}{\sqrt{\kappa_M}} + \alpha_2 R + \beta_1 \frac{r}{\sqrt{\kappa_M}} + \beta_2 R + \sqrt{\frac{M\log(M)}{n}}\frac{r}{\sqrt{\kappa_M}}\right) \\
&\times \left(1 + \sum_{t=1}^\infty \eta(t+1)e^{-t}\right).
\end{aligned}$$

Since

$$\sum_{t=1}^\infty \eta(t+1)e^{-t} \leq \int_{t=1}^\infty \left(\sqrt{t+1} + \frac{t+1}{\sqrt{n}}\right) e^{-(t-1)} \mathrm{d}t \leq 5,$$

we obtain

$$R_n(\mathcal{H}^{(r)}_\psi(R)) \leq 6\tilde{\phi}\left(\alpha_1 \frac{r}{\sqrt{\kappa_M}} + \alpha_2 R + \beta_1 \frac{r}{\sqrt{\kappa_M}} + \beta_2 R + \sqrt{\frac{M\log(M)}{n}}\frac{r}{\sqrt{\kappa_M}}\right).$$

By re-setting $\tilde{\phi} \leftarrow 6\tilde{\phi}$, we obtain the local Rademacher complexity upper bound.

Acknowledgments

We would like to thank Marius Kloft, Gilles Blanchard, Ryota Tomioka and Masashi Sugiyama for suggestive discussions. This work was partially supported by MEXT Kakenhi (25730013, 25120012, 26280009, and 15H05707), JST-PRESTO (No. JPMJPR14E4) and JST-CREST (No. JPMJCR1304 and JPMJCR14D7).

References

J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. S. Nath, and S. Raman. Variable sparsity kernel learning. *Journal of Machine Learning Research*, 12:565–592, 2011. MR2783177

A. Argyriou, R. Hauser, C. A. Micchelli, and M. Pontil. A DC-programming algorithm for kernel selection. In *the 23st International Conference on Machine Learning*, 2006.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008. MR2417268

F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 105–112. 2009.

F. R. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *the 21st International Conference on Machine Learning*, pages 41–48, 2004.

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33:1487–1537, 2005. MR2166554

P. Bartlett, M. Jordan, and D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006. MR2268032

C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988. MR0928802

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. *C. R. Acad. Sci. Paris Ser. I Math.*, 334:495–500, 2002. MR1890640

U. Chakraborty, editor. *Advances in Differential Evolution (Studies in Computational Intelligence)*. Springer, 2008.

C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 396–404. 2009a.

C. Cortes, M. Mohri, and A. Rostamizadeh. L_2 regularization for learning kernels. In *the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, 2009b. Montréal, Canada.

C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge, Cambridge, 1996. MR1410258

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015. MR3588285

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971. MR0290013

M. Kloft and G. Blanchard. The local rademacher complexity of lp-norm multiple kernel learning, 2011. arXiv:1103.0790. MR3464982

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate ℓ_p-norm multiple kernel learning. In *Advances in Neural Information Processing Systems 22*, pages 997–1005, Cambridge, MA, 2009. MIT Press. MR2786915

M. Kloft, U. Rückert, and P. L. Bartlett. A unifying view of multiple kernel learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2010.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm multiple kernel learning, 2011. MR2786915

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34:2593–2656, 2006. MR2329442

V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of the Annual Conference on Learning Theory*, pages 229–238, 2008.

V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010. MR2766864

K. P. . R. M. S. . J. A. Lampinen. *Differential Evolution - A Practical Approach to Global Optimization*. Springer, 2005. MR2191377

G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004. MR2247973

M. Ledoux and M. Talagrand. *Probability in Banach Spaces. Isoperimetry and Processes*. Springer, New York, 1991. MR1102015. MR1102015

L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009. MR2572443

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005. MR2249850

C. A. Micchelli, M. Pontil, Q. Wu, and D.-X. Zhou. Error bounds for learning the kernel. *Analysis and Applications*, 14(06):849–868, 2016. MR3564937

C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, 2005. MR2249848

G. Raskutti, M. Wainwright, and B. Yu. Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems 22*, pages 1563–1570. MIT Press, Cambridge, MA, 2009.

G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse ad-

ditive models over kernel classes via convex programming. Technical report, 2010. arXiv:1008.3654. MR2913704

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

J. Shawe-Taylor. Kernel learning for novelty detection. In *NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, Whistler, 2008.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of the Annual Conference on Learning Theory*, 2006. MR2280605

I. Steinwart. *Support Vector Machines*. Springer, 2008. MR2450103

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pages 79–93, 2009.

T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *The Annals of Statistics*, 41(3):1381–1405, 2013. MR3113815

T. Suzuki and R. Tomioka. SpicyMKL: A fast algorithm for multiple kernel learning with thousands of kernels. *Machine Learning*, 85:77–108, 2011. MR3108229

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996. MR1419006

R. Tomioka and T. Suzuki. Sparsity-accuracy trade-off in MKL. In *NIPS 2009 Workshop: Understanding Multiple Kernel Learning Methods*, Whistler, 2009.

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996. MR1385671

M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *The 26th International Conference on Machine Learning*, 2009.

Q. Wu, Y. Ying, and D.-X. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007. MR2297018

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. MR1742500

Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In S. Dasgupta and A. Klivans, editors, *Proceedings of the Annual Conference on Learning Theory*, Montreal Quebec, 2009. Omnipress.

Y. Ying and D.-X. Zhou. Learnability of gaussians with flexible variances. *Journal of Machine Learning Research*, 8(Feb):249–276, 2007. MR2320669

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006. MR2212574