Fast Learning Rate of Multiple Kernel Learning: Trade-Off between Sparsity and Smoothness

Taiji Suzuki
The University of Tokyo
s-taiji@stat.t.u-tokyo.ac.jp

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

Abstract

We investigate the learning rate of multiple kernel leaning (MKL) with ℓ_1 and elastic-net regularizations. The elastic-net regularization is a composition of an ℓ_1-regularizer for inducing the sparsity and an ℓ_2-regularizer for controlling the smoothness. We focus on a sparse setting where the total number of kernels is large but the number of non-zero components of the ground truth is relatively small, and show sharper convergence rates than the learning rates ever shown for both ℓ_1 and elastic-net regularizations. Our analysis shows there appears a trade-off between the sparsity and the smoothness when it comes to selecting which of ℓ_1 and elastic-net regularizations to use; if the ground truth is smooth, the elastic-net regularization is preferred, otherwise the ℓ_1 regularization is preferred.

1 Introduction

Learning with kernels such as support vector machines has been demonstrated to be a promising approach, given that kernels were chosen appropriately (Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004). So far, various strategies have been employed for choosing appropriate kernels, ranging from simple cross-validation (Chapelle et al., 2002) to more sophisticated "kernel learning" approaches (Ong et al., 2005, Argyriou et al., 2006, Bach, 2009, Cortes et al., 2009b, Varma and Babu, 2009).

Multiple kernel learning (MKL) is one of the systematic approaches to learning kernels, which tries to find the optimal linear combination of prefixed base-kernels by convex optimization (Lanckriet et al., 2004). The seminal paper by Bach et al. (2004) showed that this linear-combination MKL formulation can be interpreted as ℓ_1-mixed-norm regularization (i.e., the sum of the norms of the base kernels). Based on this interpretation, several variations of MKL were proposed, and promising performance was achieved by 'intermediate' regularization strategies between the sparse (ℓ_1) and dense (ℓ_2) regularizers, e.g., a mixture of ℓ_1-mixed-norm and ℓ_2-mixed-norm called the *elastic-net regularization* (Shawe-Taylor, 2008, Tomioka and Suzuki, 2009) and ℓ_p-mixed-norm regularization with $1 < p < 2$ (Micchelli and Pontil, 2005, Kloft et al., 2009).

Together with the active development of practical MKL optimization algorithms, theoretical analysis of MKL has also been extensively conducted. For ℓ_1-mixed-norm MKL, Koltchinskii and Yuan (2008) established the learning rate $d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} + d\log(M)/n$ under rather restrictive conditions, where n is the number of samples, d is the number of non-zero components of the ground truth, M is the number of kernels, and s ($0 < s < 1$) is a constant representing the complexity of the reproducing kernel Hilbert spaces (RKHSs). Their conditions include a smoothness assumption of the ground truth. For elastic-net regularization (which we call *elastic-net MKL*), Meier et al. (2009) gave a near optimal convergence rate $d\,(n/\log(M))^{-\frac{1}{1+s}}$. Recently, Koltchinskii and Yuan (2010) showed that MKL with a variant of ℓ_1-mixed-norm regularization (which we call L_1*-MKL*) achieves the minimax optimal convergence rate, which successfully captured sharper dependency with respect to $\log(M)$ than the bound of Meier et al. (2009) and established the bound $dn^{-\frac{1}{1+s}} + d\log(M)/n$. Another line of research considers the cases where the ground truth is not sparse, and bounds the Rademacher complexity of a candidate kernel class by a pseudo-dimension of the kernel class (Srebro and Ben-David, 2006, Ying and Campbell, 2009, Cortes et al., 2009a,

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22. Copyright 2012 by the authors.

Table 1: Relation between our analysis and existing analyses.

	regularizer	smoothness (q)	minimaxity	convergence rate
Koltchinskii and Yuan (2008)	ℓ_1	$q=1$	?	$d^{\frac{1-s}{1+s}}n^{-\frac{1}{1+s}}+\frac{d\log(M)}{n}$
Meier et al. (2009)	Elastic-net	$q=0$	not optimal	$d\left(\frac{\log(M)}{n}\right)^{\frac{1}{1+s}}$
Koltchinskii and Yuan (2010)	ℓ_1	$q=0$	ℓ_∞-ball	$(d+R_{1,f^*})n^{-\frac{1}{1+s}}+\frac{d\log(M)}{n}$
This paper	Elastic-net	$0\le q\le 1$	ℓ_2-ball	$\left(\frac{d}{n}\right)^{\frac{1+q}{1+q+s}}R_{2,g^*}^{\frac{2s}{1+q+s}}+\frac{d\log(M)}{n}$
	ℓ_1	$q=0$	ℓ_1-ball	$d^{\frac{1-s}{1+s}}n^{-\frac{1}{1+s}}R_{1,f^*}^{\frac{2s}{1+s}}+\frac{d\log(M)}{n}$

Kloft et al., 2010).

In this paper, we focus on the sparse setting (i.e., the total number of kernels is large, but the number of non-zero components of the ground truth is relatively small), and derive sharp learning rates for both L_1-MKL and elastic-net MKL. Our new learning rates,

$$(L_1\text{-MKL}) \qquad d^{\frac{1-s}{1+s}}n^{-\frac{1}{1+s}}R_{1,f^*}^{\frac{2s}{1+s}}+\frac{d\log(M)}{n},$$

$$(\text{Elastic-net MKL}) \quad d^{\frac{1+q}{1+q+s}}n^{-\frac{1+q}{1+q+s}}R_{2,g^*}^{\frac{2s}{1+q+s}}+\frac{d\log(M)}{n},$$

are faster than all the existing bounds, where R_{1,f^*} is the ℓ_1-mixed-norm of the truth, R_{2,g^*} is a kind of ℓ_2-mixed-norm of the truth, and q ($0\le q\le 1$) is a constant depending on the smoothness of the ground truth.

Our contributions are summarized as follows:

(a) The sharpest existing bound for L_1-MKL given by Koltchinskii and Yuan (2010) achieves the minimax rate on the ℓ_∞-mixed-norm ball (Raskutti et al., 2009, 2010). Our work follows this line and show that the learning rates for L_1-MKL and elastic-net MKL further achieve the minimax rates on the ℓ_1-*mixed-norm ball* and ℓ_2-*mixed-norm ball* respectively, both of which are faster than that on the ℓ_∞-mixed-norm ball. This result implies that the bound by Koltchinskii and Yuan (2010) is tight only when the ground truth is evenly spread in the non-zero components.

(b) We included the *smoothness* q of the ground truth into our learning rate, where the ground truth is said to be smooth if it is represented as a convolution of a certain function and an integral kernel (see Assumption 2). Intuitively, for larger q, the truth is smoother. We show that elastic-net MKL properly makes use of the smoothness of the truth: The smoother the truth is, the faster the convergence rate of elastic-net MKL is. That is, the resultant convergence rate of elastic-net MKL becomes as if the complexity of RKHSs was $\frac{s}{1+q}$ instead of the true complexity s. Meier et al. (2009) and Koltchinskii and Yuan (2010) assumed $q=0$ and Koltchinskii and Yuan (2008) considered a situation of $q=1$. Our analysis covers both of those situations, and is more general since any $0\le q\le 1$ is allowed.

(c) We show that there is a trade-off between the sparsity and the smoothness of the estimator. While L_1-MKL gives a sparser solution than elastic-net, a smoother solution is generated by elastic-net MKL. Our analysis claims that when the smoothness q of the truth is small (say $q=0$), L_1-MKL achieves a faster convergence rate than elastic-net MKL. On the other hand, if the truth is smooth, the learning rate of elastic-net MKL could be faster than L_1-MKL.

The relation between our analysis and existing analyses is summarized in Table 1.

2 Preliminaries

In this section, we formulate elastic-net MKL, and summarize mathematical tools that are needed for our theoretical analysis.

Formulation Suppose we are given n samples $\{(x_i,y_i)\}_{i=1}^n$ where x_i belongs to an input space $\mathcal{X}$ and $y_i\in\mathbb{R}$. We denote the marginal distribution of X by Π. We consider an MKL regression problem in which the unknown target function is represented as $f(x)=\sum_{m=1}^M f_m(x)$, where each f_m belongs to a different RKHS $\mathcal{H}_m$ ($m=1,\dots,M$) with kernel k_m over $\mathcal{X}\times\mathcal{X}$.

The elastic-net MKL we consider in this paper is the version considered in Meier et al. (2009):

$$\hat{f}=\mathop{\arg\min}_{\substack{f_m\in\mathcal{H}_m\\(m=1,\dots,M)}}\frac{1}{n}\sum_{i=1}^N\left(y_i-\sum_{m=1}^M f_m(x_i)\right)^2 +\sum_{m=1}^M\left(\lambda_1^{(n)}\|f_m\|_n+\lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m}+\lambda_3^{(n)}\|f_m\|_{\mathcal{H}_m}^2\right), \tag{1}$$

where $\|f_m\|_n:=\sqrt{\frac{1}{n}\sum_{i=1}^n f_m(x_i)^2}$ and $\|f_m\|_{\mathcal{H}_m}$ is the RKHS norm of f_m in $\mathcal{H}_m$. The regularizer is the mixture of ℓ_1-term $\sum_{m=1}^M(\lambda_1^{(n)}\|f_m\|_n+\lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m})$ and ℓ_2-term $\sum_{m=1}^M\lambda_3^{(n)}\|f_m\|_{\mathcal{H}_m}^2$. In that sense, we say that the regularizer is of the elastic-net type[1]

[1]There is another version of MKL with elastic-net

(Zou and Hastie, 2005). Here the ℓ_1-term is a mixture of the empirical L_2-norm $\|f_m\|_n$ and the RKHS norm $\|f_m\|_{\mathcal{H}_m}$. Koltchinskii and Yuan (2010) considered ℓ_1-regularization that contains only the ℓ_1-term: $\sum_m \lambda_1^{(n)}\|f_m\|_n + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m}$. To distinguish the situations of $\lambda_3^{(n)} = 0$ and $\lambda_3^{(n)} > 0$, we refer to the learning method (1) with $\lambda_3^{(n)} = 0$ as *L_1-MKL* and that with $\lambda_3^{(n)} > 0$ as *elastic-net MKL*.

By the representer theorem (Kimeldorf and Wahba, 1971), the solution $\hat{f}$ can be expressed as a linear combination of nM kernels: $\exists \alpha_{m,i} \in \mathbb{R}$, $\hat{f}_m(x) = \sum_{i=1}^n \alpha_{m,i} k_m(x, x_i)$. Thus, using the Gram matrix $\boldsymbol{K}_m = (k_m(x_i, x_j))_{i,j}$, the regularizer in (1) is expressed as

$$\sum_{m=1}^M \left(\lambda_1^{(n)} \sqrt{\boldsymbol{\alpha}_m^\top \frac{\boldsymbol{K}_m \boldsymbol{K}_m}{n} \boldsymbol{\alpha}_m} + \lambda_2^{(n)} \sqrt{\boldsymbol{\alpha}_m^\top \boldsymbol{K}_m \boldsymbol{\alpha}_m} + \lambda_3^{(n)} \boldsymbol{\alpha}_m^\top \boldsymbol{K}_m \boldsymbol{\alpha}_m \right),$$

where $\boldsymbol{\alpha}_m = (\alpha_{m,i})_{i=1}^n \in \mathbb{R}^n$. Thus, we can solve the problem by an SOCP (second-order cone programming) solver as in Bach et al. (2004), the coordinate descent algorithms (Meier et al., 2008) or the alternating direction method of multipliers (Boyd et al., 2011).

Notations and Assumptions Here, we present several assumptions used in our theoretical analysis and prepare notations.

Let $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_M$. We utilize the same notation $f \in \mathcal{H}$ indicating both the vector $(f_1, \ldots, f_M)$ 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 $L_2(\Pi)$. However, this will not cause any confusion. We denote by $f^* \in \mathcal{H}$ the ground truth satisfying the following assumption (the decomposition $f^* = \sum_{m=1}^M f_m^*$ of the truth might not be unique but we fix one possibility).

Assumption 1. (Basic Assumptions)

(A1-1) *There exists $f^* = (f_1^*, \ldots, f_M^*) \in \mathcal{H}$ such that $\mathrm{E}[Y|X] = \sum_{m=1}^M f_m^*(X)$, and the noise $\epsilon := Y - f^*(X)$ is bounded as $|\epsilon| \leq L$.*

(A1-2) *For each $m = 1, \ldots, M$, $\mathcal{H}_m$ is separable and $\sup_{X \in \mathcal{X}} |k_m(X, X)| \leq 1$.*

regularization considered in Shawe-Taylor (2008) and Tomioka and Suzuki (2009), that is, $\lambda_2^{(n)} \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m} + \lambda_3^{(n)} \sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^2$ (i.e., there is no $\|f_m\|_n$ term in the regularizer). However, we focus on Eq. (1) because the above one is too loose to properly bound the irrelevant components of the estimated function.

The first assumption in (A1-1) ensures the model $\mathcal{H}$ is correctly specified, and the technical assumption $|\epsilon| \leq L$ allows ϵf to be Lipschitz continuous with respect to f. These assumptions are not essential, and can be relaxed to misspecified models and unbounded noise such as Gaussian noise (Raskutti et al., 2010). However, for the sake of simplicity, we assume these conditions. It is known that the assumption (A1-2) gives the relation $\|f_m\|_\infty \leq \|f_m\|_{\mathcal{H}_m}$ (see Chapter 4 of Steinwart and Christmann (2008)).

We define an operator $T_m : \mathcal{H}_m \to \mathcal{H}_m$ as

$$\langle f_m, T_m g_m \rangle_{\mathcal{H}_m} := \mathrm{E}[f_m(X) g_m(X)],$$

where $f_m, g_m \in \mathcal{H}_m$. Due to Mercer's theorem, there are an orthonormal system $\{\phi_{\ell,m}\}_{\ell,m}$ in $L_2(\Pi)$ and the spectrum $\{\mu_{\ell,m}\}_{\ell,m}$ such that k_m has the following spectral representation:

$$k_m(x, x') = \sum_{\ell=1}^\infty \mu_{\ell,m} \phi_{\ell,m}(x) \phi_{\ell,m}(x'). \tag{2}$$

By this spectral representation, the inner product of the RKHS $\mathcal{H}_m$ can be expressed as $\langle f_m, g_m \rangle_{\mathcal{H}_m} = \sum_{\ell=1}^\infty \mu_{\ell,m}^{-1} \langle f_m, \phi_{\ell,m} \rangle_{L_2(\Pi)} \langle \phi_{\ell,m}, g_m \rangle_{L_2(\Pi)}$.

Assumption 2. (Convolution Assumption) *There exist a real number $0 \leq q \leq 1$ and $g_m^* \in \mathcal{H}_m$ such that*

$$\text{(A2)} \qquad f_m^*(x) = \int_{\mathcal{X}} k_m^{(q/2)}(x, x') g_m^*(x') \mathrm{d}\Pi(x'), \quad (\forall m = 1, \ldots, M),$$

where $k_m^{(q/2)}(x, x') = \sum_{k=1}^\infty \mu_{k,m}^{q/2} \phi_{k,m}(x) \phi_{k,m}(x')$. This is equivalent to the following operator representation:

$$f_m^* = T_m^{\frac{q}{2}} g_m^*.$$

We define $g^* \in \mathcal{H}$ as $g^* = (g_1^*, \cdots, g_M^*)$ and $g^* = \sum_{m=1}^M g_m^*$.

The constant q represents the smoothness of the truth f_m^* because f_m^* is a convolution of the integral kernel $k_m^{(q/2)}$ and g_m^*, and high frequency components are suppressed as q becomes large. Therefore, as q becomes larger, f^* becomes "smoother". The assumption (A2) was considered in Caponnetto and de Vito (2007) to analyze the convergence rate of least-squares estimators in a single kernel setting. In MKL settings, Koltchinskii and Yuan (2008) showed a fast learning rate of MKL, and Bach (2008) employed the assumption for $q = 1$ to show the consistency of MKL. Proposition 9 of Bach (2008) gave a sufficient condition to fulfill (A2) with $q = 1$ for translation invariant kernels $k_m(x, x') = h_m(x - x')$. Meier et al. (2009) considered a situation with $q = 0$ on Sobolev space; the analysis

of Koltchinskii and Yuan (2010) also corresponds to $q = 0$. Note that (A2) with $q = 0$ imposes nothing on the smoothness about the truth, and our analysis also covers this case.

We show in Appendix A (supplementary material) that as q increases, the space of the functions that satisfy (A2) becomes "simpler". Thus, it might be natural to expect that, under the Convolution Assumption (A2), the learning rate becomes faster as q increases. Although this conjecture is actually true, it is not obvious because the Convolution Assumption only restricts the ground truth, not the search space.

Next we introduce a parameter representing the complexity of RKHSs.

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

$$\text{(A3)} \qquad \mu_{j,m} \leq c j^{-\frac{1}{s}}, \quad (1 \leq \forall j, 1 \leq \forall m \leq M),$$

where $\{\mu_{j,m}\}_{j=1}^{\infty}$ is the spectrum of the kernel k_m (see Eq.(2)).

It was shown that the spectral assumption (A3) is equivalent to the classical covering number assumption[2] (Steinwart et al., 2009), but the spectral formulation gives a clearer insight for the complexity of the set of the smooth functions introduced in Assumption 2 (see Appendix A). If the spectral assumption (A3) holds, there exists a constant C that depends only on s and c such that

$$\mathcal{N}(\varepsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi)) \leq C\varepsilon^{-2s}, \quad (1 \leq \forall m \leq M), \qquad (3)$$

and the converse is also true (see Theorem 15 of Steinwart et al. (2009) and Steinwart and Christmann (2008) for details). Therefore, if s is large, at least one of the RKHSs is "complex"; and if s is small, all the RKHSs are "simple". A more detailed characterization of the covering number in terms of the spectrum is provided in Appendix A in the supplementary material. The covering number of the space of functions that satisfy the Convolution Assumption (A2) is also provided there.

We denote by I_0 the indices of truly active kernels, i.e.,

$$I_0 := \{m \mid \|f_m^*\|_{\mathcal{H}_m} > 0\}.$$

We define the number of truly active components as $d := |I_0|$. For $f = \sum_{m=1}^{M} f_m \in \mathcal{H}$ and a subset of indices $I \subseteq \{1, \ldots, M\}$, we define $\mathcal{H}_I = \oplus_{m \in I} \mathcal{H}_m$ and denote by $f_I \in \mathcal{H}_I$ the restriction of f to an index set I, i.e., $f_I = \sum_{m \in I} f_m$. For a given set of indices $I \subseteq \{1, \ldots, M\}$, we introduce a quantity $\kappa(I)$ representing the correlation of RKHSs inside the indices I:

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

Similarly, we define the *canonical correlations* of RKHSs between I and I^c as follows:

$$\rho(I) := \sup \left\{ \frac{\langle f_I, g_{I^c} \rangle_{L_2(\Pi)}}{\|f_I\|_{L_2(\Pi)} \|g_{I^c}\|_{L_2(\Pi)}} \;\middle|\; f_I \in \mathcal{H}_I, g_{I^c} \in \mathcal{H}_{I^c}, \right. \\ \left. f_I \neq 0, g_{I^c} \neq 0 \right\}.$$

These quantities give a connection between the $L_2(\Pi)$-norm of $f \in \mathcal{H}$ and the $L_2(\Pi)$-norm of $\{f_m\}_{m \in I}$ as shown in the following lemma. The proof is given in Appendix B in the supplementary material.

Lemma 1. *For all $I \subseteq \{1, \ldots, M\}$, we have*

$$\|f\|_{L_2(\Pi)}^2 \geq (1 - \rho(I)^2) \kappa(I) \left(\sum_{m \in I} \|f_m\|_{L_2(\Pi)}^2 \right). \qquad (4)$$

We impose the following assumption for $\kappa(I_0)$ and $\rho(I_0)$.

Assumption 4. (Incoherence Assumption) *For the truly active components I_0, $\kappa(I_0)$ is strictly positive and $\rho(I_0)$ is strictly less than 1:*

$$\text{(A4)} \qquad 0 < \kappa(I_0)(1 - \rho^2(I_0)).$$

This condition is known as the *incoherence condition* (Koltchinskii and Yuan, 2008, Meier et al., 2009), i.e., RKHSs are not too dependent on each other. In addition to the lower bound (4), we also obtain an upper bound of the $L_2(\Pi)$-norm of $\hat{f} - f^*$ using the $L_2(\Pi)$-norms of $\{\hat{f}_m - f_m^*\}_{m \in I_0}$. Thus, by the incoherence condition and Lemma 1, we may focus on bounding the $L_2(\Pi)$-norm of the "low-dimensional" components $\{\hat{f}_m - f_m^*\}_{m \in I_0}$, instead of all the components. Koltchinskii and Yuan (2010) considered a weaker condition including the *restricted isometry* (Candes and Tao, 2007) instead of (A4). Such a weaker condition is also applicable to our analysis, but we employ (A4) for simplicity.

Finally, we impose the following technical assumption related to the sup-norm of the members in the RKHSs.

Assumption 5. (Sup-norm Assumption) *Along with the Spectral Assumption (A3), there exists a constant C_1 such that*

$$\text{(A5)} \qquad \|f_m\|_{\infty} \leq C_1 \|f_m\|_{L_2(\Pi)}^{1-s} \|f_m\|_{\mathcal{H}_m}^{s},$$

[2] The ϵ-covering number $\mathcal{N}(\epsilon, \mathcal{B}_{\mathcal{H}_m}, L_2(\Pi))$ 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).

$$(\forall f_m \in \mathcal{H}_m, m = 1, \dots, M),$$

where s is the exponent defined in the Spectral Assumption (A3).

This assumption might look a bit strong, but this is satisfied if the RKHS is a Sobolev space or is continuously embeddable in a Sobolev space. For example, the RKHSs of Gaussian kernels are continuously embedded in all Sobolev spaces, and thus satisfy the sup-norm Assumption (A5). More generally, RKHSs with γ-times continuously differentiable kernels on a closed Euclidean ball in $\mathbb{R}^d$ are also continuously embedded in a Sobolev space, and satisfy the sup-norm Assumption (A5) with $s = \frac{d}{2\gamma}$ (see Corollary 4.36 of Steinwart and Christmann (2008)). Therefore, this assumption is common for practically used kernels. A more general necessary and sufficient condition in terms of *real interpolation* is shown in Bennett and Sharpley (1988). Steinwart et al. (2009) used this assumption to show the optimal convergence rates for regularized regression with a single kernel function where the true function is not contained in the model, and one can find detailed discussions about the assumption there.

3 Convergence Rate Analysis

In this section, we present our main result.

3.1 The Convergence Rate of L_1-MKL and Elastic-net MKL

Here we derive the learning rate of the estimator $\hat{f}$ defined by Eq. (1). We may suppose that the number of kernels M and the number of active kernels d are increasing with respect to the number of samples n. Our main purpose of this section is to show that the learning rate can be faster than the existing bounds. The existing bound has already been shown to be optimal on the ℓ_∞-mixed-norm ball (Koltchinskii and Yuan, 2010, Raskutti et al., 2010). Our claim is that the convergence rates can further achieve the minimax optimal rates on the *ℓ_1-mixed-norm ball* and *ℓ_2-mixed-norm ball*, which are faster than that on the ℓ_∞-mixed-norm ball.

Define $\eta(t)$ for $t > 0$ and $\xi_n(\lambda)$ for given $\lambda > 0$ as

$$\eta(t) := \max(1, \sqrt{t}, t/\sqrt{n}), \tag{5a}$$

$$\xi_n := \xi_n(\lambda) = \max\left(\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}}, \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}}, \sqrt{\frac{\log(M)}{n}}\right). \tag{5b}$$

For a given function $f \in \mathcal{H}$ and $1 \le p \le \infty$, we define the ℓ_p-mixed-norm of f as

$$R_{p,f} := \left(\sum_{m=1}^M \|f_m\|_{\mathcal{H}_m}^p\right)^{\frac{1}{p}}.$$

Then we obtain the convergence rate of L_1- and elastic-net MKL as follows.

Theorem 2 (Convergence Rate of L_1-MKL and Elastic-net MKL). *Suppose Assumptions 1–5 are satisfied. Then there exist constants $\tilde{C}$ and ψ_s depending on $s, c, L, C_1, \rho(I_0), \kappa(I_0)$ such that, for all n sufficiently large, the following convergence rates hold:*

(L_1-MKL) *If $\lambda_1^{(n)} = \psi_s \eta(t) \xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda_1^{(n)} \lambda^{\frac{1}{2}}$, $\lambda_3^{(n)} = 0$, where $\lambda = d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{-\frac{2}{1+s}}$, the generalization error of L_1-MKL is bounded as*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \le \tilde{C}\left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} + d^{\frac{s-1}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2}{1+s}} + \frac{d\log(M)}{n}\right)\eta(t)^2, \tag{6}$$

with high probability.

(Elastic-net MKL) *If $\lambda_1^{(n)} = \psi_s \eta(t) \xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda_1^{(n)} \lambda^{\frac{1}{2}}$, $\lambda_3^{(n)} = \lambda$, where $\lambda = d^{\frac{1}{1+q+s}} n^{-\frac{1}{1+q+s}} R_{2,g^*}^{-\frac{2}{1+q+s}}$, the generalization error of elastic-net MKL is bounded as*

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \le \tilde{C}\left(d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + d^{\frac{q+s}{1+q+s}} n^{-\frac{1+q}{1+q+s} - \frac{q(1-s)}{(1+s)(1+q+s)}} R_{2,g^*}^{\frac{2}{1+q+s}} + \frac{d\log(M)}{n}\right)\eta(t)^2, \tag{7}$$

with high probability.

The rigorous statement and the proof of Theorem 2 is provided in Appendix E in the supplementary material. The bounds presented in the theorem can be further simplified under additional weak conditions: If $R_{1,f^*} \le Cd$ with a constant C (this holds if $\|f_m^*\|_{\mathcal{H}_m} \le C$ for all m), then the first term in the learning rate (6) of L_1-MKL dominates the second term, and thus Eq. (6) becomes

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \le O_p\left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} + \frac{d\log(M)}{n}\right). \tag{8}$$

Similarly, as for the bound of elastic-net MKL, if $R_{2,g^*}^2 \le Cn^{\frac{q}{1+s}} d$ with a constant C (this holds if $\|g_m^*\|_{\mathcal{H}_m} \le \sqrt{C}$ for all m), then Eq. (7) becomes

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \le O_p\left(d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d\log(M)}{n}\right). \tag{9}$$

We note that, as s becomes smaller (the RKHSs become simpler), both learning rates of L_1-MKL and

elastic-net MKL become faster if $R_{1,f^*}, R_{2,g^*} \geq 1$. Although the solutions of both L_1-MKL and elastic-net MKL are derived from the same optimization framework (1), there appears two convergence rates (8) and (9) that posses different characteristics depending on $\lambda_3^{(n)} = 0$ or not. There appears no dependency on the smoothness parameter q in the bound (8) of L_1-MKL, while the bound (9) of elastic-net MKL depends on q. Let us compare these two learning rates on the two situations: $q = 0$ and $q > 0$.

(i) ($q = 0$) In this situation, the true function f^* is not smooth and $g^* = f^*$ from the definition of q (see Assumption 2 for the definition of g^*). The terms with respect to d are $d^{\frac{1-s}{1+s}}$ for L_1-MKL (8) and $d^{\frac{1}{1+s}}$ for elastic-net MKL (9). Thus, L_1-MKL has milder dependency on d. This might reflect the fact that L_1-MKL tends to generate sparser solutions. Moreover, one can check that the learning rate of L_1-MKL (8) is better than that of elastic-net MKL (9) because Jensen's inequality $R_{1,f^*} \leq \sqrt{d} R_{2,f^*}$ gives

$$d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} \leq d^{\frac{1}{1+s}} n^{-\frac{1}{1+s}} R_{2,f^*}^{\frac{2s}{1+s}}.$$

Therefore, when the truth is non-smooth, L_1-MKL is preferred.

(ii) ($q > 0$) We see that, as q becomes large (the truth becomes smooth), the convergence rate of elastic-net MKL becomes faster. The convergence rate with respect to n is $n^{-\frac{1+q}{1+q+s}}$ for elastic-net MKL that is faster than that of L_1-MKL ($n^{-\frac{1}{1+s}}$). This shows that elastic-net MKL properly captures the smoothness of the truth f^* using the additional ℓ_2-regularization term. As we observed above, L_1-MKL converges faster than L_2-MKL when $q = 0$. However, if f^* is sufficiently smooth (g^* is small), as q increases, there appears "phase-transition", that is, the convergence rate of elastic-net MKL turns out to be faster than that of L_1-MKL ($d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} \geq d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}}$). This indicates that, when the truth f^* is smooth, elastic-net MKL is preferred.

An interesting observation here is that depending on the smoothness q of the truth, the preferred regularization changes. This is due to the trade-off between the sparsity and the smoothness. Below, we show that these bounds (8) and (9) achieve the minimax optimal rates on the ℓ_1-mixed-norm ball and the ℓ_2-mixed-norm ball, respectively.

3.2 Minimax Learning Rate

Here we consider a simple setup to investigate the minimax rate. First, we assume that the input space $\mathcal{X}$ is expressed as $\mathcal{X} = \tilde{\mathcal{X}}^M$ for some space $\tilde{\mathcal{X}}$. Second, all the RKHSs $\{\mathcal{H}_m\}_{m=1}^M$ are the same as an RKHS $\tilde{\mathcal{H}}$ defined on $\tilde{\mathcal{X}}$. Finally, we assume that the marginal distribution Π of input is the product of a probability distribution Q, i.e., $\Pi = Q^M$. Thus, an input $x = (\tilde{x}^{(1)}, \ldots, \tilde{x}^{(M)}) \in \mathcal{X} = \tilde{\mathcal{X}}^M$ is concatenation of M random variables $\{\tilde{x}^{(m)}\}_{m=1}^M$ independently and identically distributed from the distribution Q. Moreover, the function class $\mathcal{H}$ is assumed to be a class of functions f such that $f(x) = f(\tilde{x}^{(1)}, \ldots, \tilde{x}^{(M)}) = \sum_{m=1}^M f_m(\tilde{x}^{(m)})$, where $f_m \in \tilde{\mathcal{H}}$ for all m. Without loss of generality, we may suppose that all functions in $\tilde{\mathcal{H}}$ are centered: $\mathrm{E}_{\tilde{X}\sim Q}[f(\tilde{X})] = 0$ ($\forall f \in \tilde{\mathcal{H}}$). Furthermore, we assume that the spectrum of the kernel $\tilde{k}$ corresponding to the RKHS $\tilde{\mathcal{H}}$ decays at the rate of $-\frac{1}{s}$. That is, in addition to Assumption 3, we impose the following lower bound on the spectrum: There exist c', c (> 0) such that

$$c' j^{-\frac{1}{s}} \leq \mu_j \leq c j^{-\frac{1}{s}}, \tag{10}$$

where $\{\mu_j\}_j$ is the spectrum of the kernel $\tilde{k}$ (see Eq.(2)). We also assume that the noise $\{\epsilon_i\}_{i=1}^n$ is generated by the Gaussian distribution with mean 0 and standard deviation σ.

Let $\mathcal{H}_0(d)$ be the set of functions with d non-zero components in $\mathcal{H}$ defined by $\mathcal{H}_0(d) := \{(f_1, \ldots, f_M) \in \mathcal{H} \mid \#\{m \mid \|f_m\|_{\mathcal{H}_m} \neq 0\} \leq d\}$, where $\#$ denotes the cardinality of the set. We define the ℓ_p-mixed-norm ball ($p \geq 1$) with radius R in $\mathcal{H}_0(d)$ as

$$\mathcal{H}_{\ell_p}^{d,q}(R) := \Bigg\{ f = \sum_{m=1}^M f_m \;\Big|\; \exists (g_1, \ldots, g_M) \in \mathcal{H}_0(d), \; f_m = T_m^{\frac{q}{2}} g_m, \; \Big(\sum_{m=1}^M \|g_m\|_{\mathcal{H}_m}^p\Big)^{\frac{1}{p}} \leq R \Bigg\}.$$

In Raskutti et al. (2010), the minimax learning rate on $\mathcal{H}_{\ell_\infty}^{d,0}(R)$ (i.e., $p = \infty$ and $q = 0$) was derived[3]. We show (a lower bound of) the minimax learning rate for more general settings ($1 \leq p \leq \infty$ and $0 \leq q \leq 1$) in the following theorem.

Theorem 3. *Let $\tilde{s} = \frac{s}{1+q}$. Assume $d \leq M/4$. Then the minimax learning rates are lower bounded as follows. If the radius of the ℓ_p-mixed-norm ball R_p satisfies $R_p \geq d^{\frac{1}{p}} \sqrt{\frac{\log(M/d)}{n}}$, there exists a constant $\tilde{C}_1$ such that*

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_p}^{d,q}(R_p)} \mathrm{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] \geq \tilde{C}_1 \left(d^{1 - \frac{2\tilde{s}}{p(1+\tilde{s})}} n^{-\frac{1}{1+\tilde{s}}} R_p^{\frac{2\tilde{s}}{1+\tilde{s}}} + \frac{d \log(M/d)}{n} \right), \tag{11}$$

where 'inf' is taken over all measurable functions of the samples $\{(x_i, y_i)\}_{i=1}^n$ and the expectation is taken for the sample distribution.

[3]The set $\mathcal{F}_{M,d,\mathcal{H}}(R)$ in Raskutti et al. (2010) corresponds to $\mathcal{H}_{\ell_\infty}^{d,0}(R)$ in the current paper.

A proof of Theorem 3 is provided in Appendix H in the supplementary material.

Substituting $q = 0$ and $p = 1$ into the minimax learning rate (11), we see that the learning rate (8) of L_1-MKL achieves the minimax optimal rate of the ℓ_1-mixed-norm ball for $q = 0$. Moreover, the learning rate of L_1-MKL (that is minimax optimal on the ℓ_1-mixed-norm ball) is *fastest* among all the optimal minimax rates on ℓ_p-mixed-norm ball for $p \geq 1$ when $q = 0$. To see this, let $R_{p,f^*} := \left(\sum_m \|f_m^*\|_{\mathcal{H}_m}^p\right)^{\frac{1}{p}}$; then we always have $R_{1,f^*} \leq d^{1-\frac{1}{p}} R_{p,f^*} \leq d R_{\infty,f^*}$ due to Jensen's inequality, and consequently we have

$$d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} \leq d^{1-\frac{2s}{p(1+s)}} n^{-\frac{1}{1+s}} R_{p,f^*}^{\frac{2s}{1+s}} \leq d n^{-\frac{1}{1+s}} R_{\infty,f^*}^{\frac{2s}{1+s}}. \quad (12)$$

On the other hand, the learning rate (9) of elastic-net MKL achieves the minimax optimal rate (11) on the ℓ_2-mixed-norm ball ($p = 2$). When $q = 0$, the rate of elastic-net MKL is slower than that of L_1-MKL, but the optimal rate is achieved over the whole range of smoothness parameter $0 \leq q \leq 1$, which is advantageous against L_1-MKL. Moreover, the optimal rate on the ℓ_2-mixed-norm ball is still faster than that on the ℓ_∞-mixed-norm ball due to the relation (12).

The learning rates of both L_1 and elastic-net MKL coincide with the minimax optimal rate of the ℓ_∞-mixed-norm ball when the truth is *homogeneous*. For simplicity, assume $q = 0$. If $\|f_m^*\|_{\mathcal{H}_m} = 1$ ($\forall m \in I_0$) and $f_m^* = 0$ (otherwise), then $R_{p,f^*} = d^{\frac{1}{p}}$. Thus, both rates are $d n^{-\frac{1}{1+s}} + \frac{d \log(M)}{n}$ that is the minimax rate on the ℓ_∞-mixed-norm ball. We also notice that this homogeneous situation is the only situation where those convergence rates coincide with each other. As seen later, the existing bounds are the minimax rate on the ℓ_∞-mixed-norm ball, and thus are tight only in the homogeneous setting.

3.3 Optimal Parameter Selection

We need the knowledge of parameters such as $q, s, d, R_{1,f^*}, R_{2,g^*}$ to obtain the optimal learning rate shown in Theorem 2. However this is not realistic in practice.

To overcome this problem, we give an algorithmic procedure such as *cross-validation* to achieve the optimal learning rate. Roughly speaking, we split the data into the training set and the validation set and utilize the validation set to choose the optimal parameter. Given the data $D = \{(x_i, y_i)\}_{i=1}^n$, the training set D_{tr} is generated by using the half of the given data $D_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{n'}$ where $n' = \lfloor \frac{n}{2} \rfloor$ and the remaining data is used as the validation set $D_{\mathrm{te}} = \{(x_i, y_i)\}_{i=n'+1}^n$. Let $\hat{f}_\Lambda$ be the estimator given by our MKL formulation (1) where the parameter setting $\Lambda = (\lambda_1^{(n)}, \lambda_2^{(n)}, \lambda_3^{(n)})$ is employed and the training set D_{tr} is used instead of the whole data set D.

We utilize a *clipped estimator* to let the estimator bounded so that the validation procedure is effective. Given the estimator $\hat{f}_\Lambda$ and a positive real $B > 0$, the clipped estimator $\tilde{f}_\Lambda$ is given as

$$\tilde{f}_\Lambda(x) := \begin{cases} B & (B \leq \hat{f}_\Lambda(x)), \\ \hat{f}_\Lambda(x) & (-B < \hat{f}_\Lambda(x) < B), \\ -B & (\hat{f}_\Lambda(x) \leq -B). \end{cases}$$

To appropriately choose B, we assume that we can roughly estimate the range of y, and B is set to satisfy $|y| < B$ almost surely. This assumption is not unrealistic because if we set B sufficiently large so that we have $\max_i |y_i| < B$, then with high probability such B satisfies $|y| < B$ (a.s.). Instead of estimating the range of y, we can set B as $\|f_m^*\|_\infty + L \leq B$ because $\|f_m^*\|_\infty + L$ bounds the range of y from above (see Assumption 1 for the definition of L). For simplicity, we assume that B is greater than (but proportional to) $\|f_m^*\|_\infty + L$. It should be noted that the clipped estimator does not make the generalization error worse:

$$\|\tilde{f}_\Lambda - f^*\|_{L_2(\Pi)} \leq \|\hat{f}_\Lambda - f^*\|_{L_2(\Pi)},$$

because $|\tilde{f}_\Lambda(x) - f^*(x)| \leq |\hat{f}_\Lambda(x) - f^*(x)|$ for all $x \in \mathcal{X}$.

Now, for a finite set of parameter candidates $\Theta_n \subset \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$, we choose an optimal parameter that minimizes the error on the validation set:

$$\Lambda_{D_{\mathrm{te}}} := \operatorname*{argmin}_{\Lambda \in \Theta_n} \frac{1}{|D_{\mathrm{te}}|} \sum_{(x_i, y_i) \in D_{\mathrm{te}}} (\tilde{f}_\Lambda(x_i) - y_i)^2. \quad (13)$$

Then we can show that the estimator $\tilde{f}_{\Lambda_{D_{\mathrm{te}}}}$ achieves the optimal learning rate. To show that, we determine the finite set Θ_n of the candidate parameters as follows: Let $\Gamma_n := \{1/n^2, 2/n^2, \ldots, 1\}$ and

$$\Theta_n = \{(\lambda_1, \lambda_2, \lambda_3) \mid \lambda_1, \lambda_3 \in \Gamma_n, \lambda_2 = \lambda_1 \lambda_3^{\frac{1}{2}}\} \cup \{(\lambda_1, \lambda_2, \lambda_3) \mid \lambda_1, \lambda \in \Gamma_n, \lambda_2 = \lambda_1 \lambda^{\frac{1}{2}}, \lambda_3 = 0\}.$$

With this parameter set, we have the following theorem that shows the optimality of the validation procedure (13).

Theorem 4. *Assume $R_{1,f^*}, R_{2,g^*} \geq 1$ and $\|f_m^*\|_{\mathcal{H}_m}, \|g_m^*\|_{\mathcal{H}_m} \leq C$ with some constant C, then under the same settings as Theorem 2, we have*

$$\|\tilde{f}_{\Lambda_{D_{\mathrm{te}}}} - f^*\|_{L_2(\Pi)}^2$$

$$\leq O_p\left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} \wedge d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d\log(M)}{n} + \frac{B^2\log(1+n)}{n}\right),$$

where $a \wedge b$ *means* $\min\{a, b\}$.

This can be shown by combining our bound in Theorem 2 and the technique used in Theorem 7.2 of Steinwart and Christmann (2008). According to Theorem 4, the estimator $\tilde{f}_{\Lambda_{D_{\mathrm{te}}}}$ with the validated parameter $\Lambda_{D_{\mathrm{te}}}$ achieves the minimum learning rate among the oracle bound for L_1-MKL (8) and that for elastic-net MKL (9) if B is sufficiently small. Therefore, our bound is almost attainable (at the cost of the term $\frac{B^2\log(1+n)}{n}$) by a simple executable algorithm.

3.4 Comparison with Existing Bounds

Finally, we compare out bound with the existing bounds. Roughly speaking, the difference from the existing bounds is summarized in the following two points (see also Table 1 summarizing the relations between our analysis and existing analyses):

(a) Our learning rate achieves the minimax rates of the ℓ_1-mixed-norm ball or the ℓ_2-mixed-norm ball, instead of the ℓ_∞-mixed-norm ball.

(b) Our bound includes the smoothing parameter q (Assumption 2), and thus is more general and faster than existing bounds.

The first bound on the convergence rate of MKL was derived by Koltchinskii and Yuan (2008), which assumed $q = 1$ and $\frac{1}{d}\sum_{m\in I_0}(\|g_m^*\|_{\mathcal{H}_m}^2/\|f_m^*\|_{\mathcal{H}_m}^2) \leq C$. Under these rather strong conditions, they showed the bound $d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} + \frac{d\log(M)}{n}$. Our convergence rate of L_1-MKL achieves this learning rate *without* the two strong conditions. Moreover, for the smooth case $q = 1$, we have shown that elastic-net MKL has a faster rate $n^{-\frac{2}{2+s}}$ instead of $n^{-\frac{1}{1+s}}$ with respect to n.

The second bound was given by Meier et al. (2009), which showed $d\,(\log(M)/n)^{\frac{1}{1+s}}$ for elastic-net regularization under $q = 0$. Their bound almost achieves the minimax rate on the ℓ_∞-mixed-norm ball except the additional $\log(M)$ term. Compared with our bound, their bound has the $\log(M)$ term and the rate with respect to d is larger than $d^{\frac{1}{1+s}}$ in our learning rate of elastic-net MKL. Moreover, our result for elastic-net MKL covers all $0 \leq q \leq 1$.

Most recently, Koltchinskii and Yuan (2010) presented the bound $n^{-\frac{1}{1+s}}(d + \sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}) + \frac{d\log(M)}{n}$ for L_1-MKL and $q = 0$. Their bound achieves the minimax rate on the ℓ_∞-mixed-norm ball, but is not as tight as our bound (8) of L_1-MKL because by Young's inequality we always have

$$d^{\frac{1-s}{1+s}}(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m})^{\frac{2s}{1+s}} \leq d + \sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}.$$

However, their bound is $d^{\frac{2s}{1+s}}$ times slower than ours if the ground truth is *inhomogeneous*. For example, when $\|f_m^*\|_{\mathcal{H}_m} = m^{-1}$ $(m \in I_0 = \{1, \ldots, d\})$ and $f_m^* = 0$ (otherwise), their bound is $n^{-\frac{1}{1+s}} d + \frac{d\log(M)}{n}$, while our bound for L_1-MKL is $n^{-\frac{1}{1+s}} d^{\frac{1-s}{1+s}} + \frac{d\log(M)}{n}$. Moreover they assumed the *global boundedness*, that is, the sup-norm of f^* is bounded by a constant: $\|f^*\|_\infty = \|\sum_{m=1}^M f_m^*\|_\infty \leq C$. This assumption is standard and does not affect the convergence rate in single kernel learning settings. However, in MKL settings, it is pointed out that the rate is not minimax optimal in large d regime (in particular $d = \Omega(\sqrt{n})$) under the global boundedness (Raskutti et al., 2010). Our analysis omits the global boundedness by utilizing the Sup-norm Assumption (Assumption 5).

All the bounds explained above focused on either $q = 0$ or 1. On the other hand, our analysis is more general in that the whole range of $0 \leq q \leq 1$ is allowed.

4 Conclusion

We presented a new learning rate of both L_1-MKL and elastic-net MKL, which is faster than the existing bounds of several MKL formulations. According to our bound, the learning rates of L_1-MKL and elastic-net MKL achieve the minimax optimal rates on the ℓ_1-mixed-norm ball and the ℓ_2-mixed-norm ball respectively, instead of the ℓ_∞-mixed-norm ball. We also showed that a procedure like cross validation gives the optimal choice of the parameters. We observed that, depending on the smoothness of the ground truth, preferred methods (L_1-MKL or elastic-net MKL) change. This theoretical insight supports the recent experimental results (Cortes et al., 2009a, Kloft et al., 2009, Tomioka and Suzuki, 2009) such that intermediate regularization between ℓ_1 and ℓ_2 often shows favorable performances.

Acknowledgement

We would like to thank Ryota Tomioka, Alexandre B. Tsybakov, and Martin Wainwright for suggestive discussions. TS was partially supported by MEXT Kakenhi 22700289 and the Aihara Project, the FIRST program from JSPS, initiated by CSTP. MS was supported by MEXT Kakenhi 23120004.

References

A. Argyriou, R. Hauser, C. A. Micchelli, and M. Pontil. A DC-programming algorithm for kernel selection. In *ICML*, pages 41–48, 2006.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *JMLR*, 9:1179–1225, 2008.

F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *NIPS*, pages 105–112. MIT Press, Cambridge, MA, 2009.

F. R. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, pages 41–48, 2004.

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

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 334:495–500, 2002.

S. Boyd, N. Parikh, E. Chu, and B. Peleato. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger. *The Annals of Statistics*, 35(6):2313–2351, 2007.

A. Caponnetto and E. de Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1/3):131, 2002.

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

C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In *NIPS*, pages 396–404. MIT Press, Cambridge, MA, 2009b.

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

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate ℓ_p-norm multiple kernel learning. In *NIPS*, pages 997–1005, Cambridge, MA, 2009. MIT Press.

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.

V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *COLT*, pages 229–238, 2008.

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

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

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

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71, 2008.

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

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *JMLR*, 6:1099–1125, 2005.

C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *JMLR*, 6:1043–1071, 2005.

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

G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical report, 2010. arXiv:1008.3654.

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 *COLT*, pages 169–183, 2006.

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

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

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

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

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

M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *ICML*, pages 1065–1072, 2009.

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

Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In *COLT*, 2009.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical: Series B*, 67(2):301–320, 2005.

—Appendix–Supplementary Material— Fast Learning Rate of Multiple Kernel Learning: Trade-Off between Sparsity and Smoothness

A Covering Number

Here, we give a detailed characterization of the covering number in terms of the spectrum using the operator T_m. Accordingly, we give the complexity of the set of functions satisfying the Convolution Assumption (Assumption 2). We extend the domain and the range of the operator T_m to the whole space of $L_2(\Pi)$, and define its power $T_m^\beta : L_2(\Pi) \to L_2(\Pi)$ for $\beta \in [0,1]$ as

$$T_m^\beta f := \sum_{k=1}^\infty \mu_{k,m}^\beta \langle f, \phi_{k,m}\rangle_{L_2(\Pi)} \phi_{k,m}, \quad (f \in L_2(\Pi)).$$

Moreover, we define a Hilbert space $\mathcal{H}_{m,\beta}$ as

$$\mathcal{H}_{m,\beta} := \{\textstyle\sum_{k=1}^\infty b_k \phi_{k,m} \mid \sum_{k=1}^\infty \mu_{k,m}^{-\beta} b_k^2 \le \infty\},$$

and equip this space with the Hilbert space norm $\|\sum_{k=1}^\infty b_k\phi_{k,m}\|_{\mathcal{H}_{m,\beta}} := \sqrt{\sum_{k=1}^\infty \mu_{k,m}^{-\beta} b_k^2}$. One can check that $\mathcal{H}_{m,1} = \mathcal{H}_m$. Here we define, for $R > 0$,

$$\mathcal{H}_m^q(R) := \{f_m = T_m^{\frac{q}{2}} g_m \mid g_m \in \mathcal{H}_m,\ \|g_m\|_{\mathcal{H}_m} \le R\}. \tag{14}$$

Then we obtain the following lemma.

Lemma 5. *$\mathcal{H}_m^q(1)$ is equivalent to the unit ball of $\mathcal{H}_{m,1+q}$: $\mathcal{H}_m^q(1) = \{f_m \in \mathcal{H}_{m,1+q} \mid \|f_m\|_{\mathcal{H}_{m,1+q}} \le 1\}$.*

This can be shown as follows. For all $f_m \in \mathcal{H}_m^q(1)$, there exists $g_m \in \mathcal{H}_m$ such that $f_m = T_m^{\frac{q}{2}} g_m$ and $\|g_m\|_{\mathcal{H}_m} \le 1$. Thus, $g_m = (T_m^{\frac{q}{2}})^{-1} f_m = \sum_{k=1}^\infty \mu_{k,m}^{-\frac{q}{2}} \langle f, \phi_{k,m}\rangle_{L_2(\Pi)} \phi_{k,m}$ and $1 \ge \|g_m\|_{\mathcal{H}_m} = \sum_{k=1}^\infty \mu_{k,m}^{-1} \langle g, \phi_{k,m}\rangle_{L_2(\Pi)}^2 = \sum_{k=1}^\infty \mu_{k,m}^{-(1+q)} \langle f, \phi_{k,m}\rangle_{L_2(\Pi)}^2$. Therefore, $f \in \mathcal{H}_m$ is in $\mathcal{H}_m^q(1)$ if and only if the norm of f in $\mathcal{H}_{m,1+q}$ is well-defined and not greater than 1.

Now Theorem 15 of Steinwart et al. (2009) gives an upper bound of the covering number of the unit ball $\mathcal{B}_{\mathcal{H}_{m,\beta}}$ in $\mathcal{H}_{m,\beta}$ as $\log \mathcal{N}(\varepsilon, \mathcal{B}_{\mathcal{H}_{m,\beta}}, L_2(\Pi)) \le C\varepsilon^{-2\frac{s}{\beta}}$, where C is a constant depending on c, s, β. This inequality with $\beta = 1$ corresponds to Eq. (3). Moreover, substituting $\beta = 1 + q$ into the above equation, we have

$$\mathcal{N}(\varepsilon, \mathcal{H}_m^q(1), L_2(\Pi)) \le C\varepsilon^{-2\frac{s}{1+q}}. \tag{15}$$

B Proof of Lemma 1

Proof. **(Lemma 1)** For $J = I^c$, we have

$$\begin{aligned} Pf^2 &= \|f_I\|_{L_2(\Pi)}^2 + 2\langle f_I, f_J\rangle_{L_2(\Pi)} + \|f_J\|_{L_2(\Pi)}^2 \ge \|f_I\|_{L_2(\Pi)}^2 - 2\rho(I)\|f_I\|_{L_2(\Pi)}\|f_J\|_{L_2(\Pi)} + \|f_J\|_{L_2(\Pi)}^2 \\ &\ge (1-\rho(I)^2)\|f_I\|_{L_2(\Pi)}^2 \ge (1-\rho(I)^2)\kappa(I)\left(\sum_{m\in I}\|f_m\|_{L_2(\Pi)}^2\right), \end{aligned} \tag{16}$$

where we used Cauchy-Schwarz's inequality in the last line. □

C Useful Inequalities

Here we describe some inequalities that are used in the proofs many times.

Young's inequality: for all $a, b \in \mathbb{R}$ and all $\alpha \in [0,1]$, we have

$$a^\alpha b^{1-\alpha} \le \alpha a + (1-\alpha) b.$$

Hölder's inequality: for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^M$ and all $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$, we have

$$\boldsymbol{a}^\top \boldsymbol{b} \leq \|\boldsymbol{a}\|_{\ell_p} \|\boldsymbol{a}\|_{\ell_q}$$

where $\|\boldsymbol{a}'\|_{\ell_p}$ is the ℓ_p-norm of the vector $\boldsymbol{a}'$: $\|\boldsymbol{a}'\|_{\ell_p} = (\sum_{m=1}^M |a_m|^p)^{\frac{1}{p}}$ for $(1 \leq p < \infty)$ and $\|\boldsymbol{a}'\|_{\ell_\infty} = \max_m\{|a_m|\}$. The special case of Hölder's inequality for $p = q = 2$ is the Cauchy-Schwarz inequality.

D Talagrand's Concentration Inequality

The following proposition is a key tool for our analysis.

Proposition 6. (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*

$$P\left(Z \geq K\left[\mathrm{E}[Z] + \sqrt{\frac{Bt}{n}} + \frac{Ut}{n}\right]\right) \leq e^{-t}. \tag{17}$$

E Proof of Theorem 2

To prove Theorem 2, we start from the following relation that is derived from the fact that $\hat{f}$ minimizes the objective function (1):

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - y_i)^2 + \sum_{m=1}^M (\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}^2) \\
\leq &\frac{1}{n}\sum_{i=1}^n (f^*(x_i) - y_i)^2 + \sum_{m\in I_0} (\lambda_1^{(n)}\|f_m^*\|_n + \lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|f_m^*\|_{\mathcal{H}_m}^2).
\end{aligned} \tag{18}$$

Through a simple calculation, we obatin

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m=1}^M (\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}^2) \\
\leq &\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 - \|\hat{f} - f^*\|_n^2\right) + \frac{1}{n}\sum_{n=1}^n\sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \\
&+ \sum_{m\in I_0} (\lambda_1^{(n)}\|f_m^*\|_n + \lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|f_m^*\|_{\mathcal{H}_m}^2).
\end{aligned}$$

To bound the right hand side, we will show the following two bounds for the first two terms (Theorems 7 and 8):

$$\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i))\right| \leq O_p\left[\xi_n(\lambda)\left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}\right)\right], \tag{19}$$

$$\begin{aligned}
&\left|\left\|\textstyle\sum_{m=1}^M (f_m^* - \hat{f}_m)\right\|_n^2 - \left\|\textstyle\sum_{m=1}^M (f_m^* - \hat{f}_m)\right\|_{L_2(\Pi)}^2\right| \\
&\qquad \leq o_p\left\{\sqrt{n}\xi_n(\lambda)^2\left[\sum_{m=1}^M \left(\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}\right)\right]^2\right\},
\end{aligned} \tag{20}$$

for an arbitrary fixed $\lambda > 0$. Substituting these relations in to Eq. (18) yields the following inequality:

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m=1}^M (\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}^2)$$

$$\leq o_p\left\{\sqrt{n}\xi_n(\lambda)^2\left[\sum_{m=1}^M\left(\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}\right)\right]^2\right\}$$
$$+O_p\left\{\sum_{m=1}^M\xi_n(\lambda)(\|\hat{f}_m-f_m^*\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|\hat{f}_m-f_m^*\|_{\mathcal{H}_m})\right\}$$
$$+\sum_{m\in I_0}(\lambda_1^{(n)}\|f_m^*\|_n+\lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m}+\lambda_3^{(n)}\|f_m^*\|_{\mathcal{H}_m}^2).$$

This is our start point. We show the convergence rates of both elastic-net MKL and L_1-MKL from this type of inequality. Thus we require inequalities like Eq. (19) and Eq. (20).

Remind the definition of $\eta(t)$:

$$\eta(t):=\max(1,\sqrt{t},t/\sqrt{n}). \tag{21}$$

We define

$$\phi_s:=\max\left\{2KL(C_s+1+C_1),K\left[8K(C_s+1+C_1)+C_1+C_1^2\right],1\right\}, \tag{22}$$

where K is the universal constant appeared in Talagrand's concentration inequality (Proposition 6). and C_s is a constant depending on s and C that will be given in Lemma 15. Moreover we define

$$\zeta_n(r,\lambda):=\min\left(\frac{r^2\log(M)}{n\xi_n(\lambda)^4\phi_s^2},\frac{r}{\xi_n(\lambda)^2\phi_s}\right).$$

Finally we introduce two events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(r)$ for given $\lambda>0$ as

$$\mathscr{E}_1(t)=\left\{\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i f_m(x_i)\right|\leq\eta(t)\phi_s\xi_n\left(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}\right),\ \forall f_m\in\mathcal{H}_m,\forall m=1,\dots,M\right\},$$
$$\mathscr{E}_2(r)=\left\{\left|\left\|\sum_{m=1}^M f_m\right\|_n^2-\left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2\right|\leq\max(\phi_s\sqrt{n}\xi_n^2,r)\left[\sum_{m=1}^M\left(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}\right)\right]^2,\right.$$
$$\left.\forall f_m\in\mathcal{H}_m,\forall m=1,\dots,M\right\}.$$

Then the following Theorems 7 and 8 indicate that the events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(r)$ hold with probabilities greater than $1-\exp(-t)$ and $1-\exp(-\zeta_n(r,\lambda))$ respectively if $\frac{\log(M)}{\sqrt{n}}\leq 1$. Thus substituting $\hat{f}_m-f_m^*$ into f_m in the definition of $\mathscr{E}_1(t)$ and $\mathscr{E}_2(r)$, we obtain Eq. (19) and Eq. (20).

Theorem 7. *Under the Basic Assumption, the Spectral Assumption and the Supnorm Assumption, when* $\frac{\log(M)}{\sqrt{n}}\leq 1$, *we have for all* $\lambda>0$ *and all* $t\geq 1$

$$P(\mathscr{E}_1(t))\geq 1-\exp(-t).$$

Theorem 8. *Under the Spectral Assumption and the Supnorm Assumption, when* $\frac{\log(M)}{\sqrt{n}}\leq 1$, *for all* $\lambda>0$ *and all* $r>0$ *we have*

$$P(\mathscr{E}_2(r))\geq 1-\exp(-\zeta_n(r,\lambda)).$$

The proofs of these two theorems will be given in Appendix F.

Next we give a bound of irrelevant components ($m\in I_0^c$) of $\hat{f}$ (that should vanish or neglectably small) in terms of the relevant components ($\hat{f}_m,f_m^*$ for $m\in I_0$) in Lemma 9. Using Theorems 7, 8 and Lemma 9, we can show the convergence rate of elastic-net MKL and L_1-MKL; the rate of elastic-net MKL will be shown in Theorem 10 and Corollary 11 and that of L_1-MKL will be shown in Theorem 12 and Corollary 13. To prove the convergence rates, we first give upper bounds of the generalization errors that depend on the choice of regularization parameters

$\lambda_1^{(n)}, \lambda_2^{(n)}$, and $\lambda_3^{(n)}$ in Theorems 10 and 12. Then we substitute optimal regularization parameters into these results in Corollaries 11 and 13. The assertion of Theorem 2 is directly obtained from Corollaries 11 and 13.

Now we show the statement of Lemma 9 that bounds irrelevant components ($m \in I_0^c$) of $\hat{f}$ in terms of the relevant components ($\hat{f}_m, f_m^*$ for $m \in I_0$).

Lemma 9. *Set $\lambda_1^{(n)} = 4\phi_s \eta(t) \xi_n(\lambda)$ and $\lambda_2^{(n)} = \lambda^{\frac{1}{2}} \lambda_1^{(n)}$ for arbitrary $\lambda > 0$ and set $\lambda_3^{(n)} > 0$ be an arbitrary positive. Then for all n and r satisfying $\frac{\log(M)}{\sqrt{n}} \leq 1$ and $\max(\phi_s \sqrt{n} \xi_n^2, r) \leq \frac{1}{8}$, we have*

$$\begin{aligned}
&\sum_{m=1}^{M} (\lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}) \\
\leq & 8 \sum_{m \in I_0} \Bigg\{ \lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} + \\
& \qquad \lambda_3^{(n)^{\frac{1+q}{2}}} \|g_m^*\|_{\mathcal{H}_m} \left(\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_3^{(n)^{\frac{1}{2}}} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} \right) \Bigg\}, \qquad (23)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{m=1}^{M} \left(\lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} \right) \\
&\leq \sum_{m \in I_0} \left(8\lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + 8\lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + 4\lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right), \qquad (24)
\end{aligned}$$

on the events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(r)$.

The proof will be given in Appendix G.

We are ready to show the convergence rates of elastic-net MKL and L_1-MKL. The following Theorem 10 and Corollary 11 gives the convergence rate of elastic-net MKL (the rate of L_1-MKL will be shown in Theorem 12 and Corollary 13).

Theorem 10. *Suppose Assumptions 1–5 are satisfied. Let $\lambda_1^{(n)} = 4\phi_s \eta(t) \xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda^{\frac{1}{2}} \lambda_1^{(n)}$, $\lambda_3^{(n)} = \lambda$ for arbitrary $\lambda > 0$. In this setting, for all n and r satisfying $\frac{\log(M)}{\sqrt{n}} \leq 1$ and*

$$\frac{256 \max(\phi_s \sqrt{n} \xi_n^2, r) \left(d + \frac{\lambda_3^{(n)^{1+q}}}{\lambda_1^{(n)^2}} \sum_{m=1}^{M} \|g_m^*\|_{\mathcal{H}_m}^2 \right)}{(1 - \rho(I_0)^2) \kappa(I_0)} \leq \frac{1}{8}, \qquad (25)$$

we have

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{104}{(1 - \rho(I_0))^2 \kappa(I_0)} \left(d \lambda_1^{(n)^2} + \lambda_3^{(n)^{1+q}} \sum_{m=1}^{M} \|g_m^*\|_{\mathcal{H}_m}^2 \right), \qquad (26)$$

with probability $1 - \exp(-t) - \exp(-\zeta_n(r, \lambda))$ for all $t \geq 1$.

Proof. **(Theorem 10)** Notice that the assumption (25) implies $\max(\phi_s \sqrt{n} \xi_n^2, r) \leq \frac{1}{8}$. Thus the condition in Lemma 9 is met. In the proof of Lemma 9, we will show that the following inequality holds on the events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(r)$ (Eq. (60)):

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \frac{1}{2} \sum_{m \in I_0^c} (\lambda_1^{(n)} \|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}) + \sum_{m \in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \\
\leq & (\|\hat{f} - f^*\|_{L_2(\Pi)}^2 - \|\hat{f} - f^*\|_n^2) + \sum_{m=1}^{M} \eta(t) \phi_s \xi_n (\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})
\end{aligned}$$

$$+\sum_{m\in I_0}\left[\frac{3}{2}(\lambda_1^{(n)}\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_2^{(n)}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m})+2\lambda_3^{(n)}\langle f_m^*, f_m^*-\hat{f}_m\rangle_{\mathcal{H}_m}\right].$$

Here on the event $\mathscr{E}_2(r)$, the above inequality gives

$$\begin{aligned}
&\|\hat{f}-f^*\|_{L_2(\Pi)}^2+\frac{1}{2}\sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_{L_2(\Pi)}+\lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m})+\sum_{m\in I_0}\lambda_3^{(n)}\|\hat{f}_m-f_m^*\|_{\mathcal{H}_m}^2\\
\leq&\underbrace{\max(\phi_s\sqrt{n}\xi_n^2,r)\left(\sum_{m=1}^M(\|\hat{f}_m-f_m^*\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|\hat{f}_m-f_m^*\|_{\mathcal{H}_m})\right)^2}_{\text{(i)}}\\
&+\underbrace{\sum_{m=1}^M\eta(t)\phi_s\xi_n(\|\hat{f}_m-f_m^*\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|\hat{f}_m-f_m^*\|_{\mathcal{H}_m})}_{\text{(ii)}}\\
&+\sum_{m\in I_0}\left[\underbrace{\frac{3}{2}(\lambda_1^{(n)}\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_2^{(n)}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m})}_{\text{(iii)}}+\underbrace{2\lambda_3^{(n)}\langle f_m^*, f_m^*-\hat{f}_m\rangle_{\mathcal{H}_m}}_{\text{(iv)}}\right].
\end{aligned}\tag{27}$$

From now on, we bound the terms (i) to (iv) in the RHS. By the assumption, we have $\lambda_3^{(n)\frac{1}{2}}=\lambda^{\frac{1}{2}}=\lambda_2^{(n)}/\lambda_1^{(n)}$. This and Eq. (23) yield

$$\begin{aligned}
&\sum_{m=1}^M(\lambda_1^{(n)}\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_2^{(n)}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m})\\
\leq&8\sum_{m\in I_0}\left(\lambda_1^{(n)}\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_2^{(n)}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}+\right.\\
&\qquad\left.\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}\left(\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_3^{(n)\frac{1}{2}}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}\right)\right)\\
=&8\sum_{m\in I_0}\left(1+\frac{\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}}\right)\left(\lambda_1^{(n)}\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_2^{(n)}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}\right).
\end{aligned}\tag{28}$$

Step 1. (Bound of the first term (i) in the RHS of Eq. (27)) By Eq. (28) and $\lambda^{\frac{1}{2}}=\lambda_3^{(n)\frac{1}{2}}=\lambda_2^{(n)}/\lambda_1^{(n)}$, the term (i) on the RHS of Eq. (27) can be upper bounded as

$$\begin{aligned}
&\max(\phi_s\sqrt{n}\xi_n^2,r)\left(\sum_{m=1}^M(\|\hat{f}_m-f_m^*\|_{L_2(\Pi)}+\lambda_3^{(n)\frac{1}{2}}\|\hat{f}_m-f_m^*\|_{\mathcal{H}_m})\right)^2\\
\leq&\max(\phi_s\sqrt{n}\xi_n^2,r)\left(8\sum_{m\in I_0}\left(1+\frac{\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}}\right)\left(\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_3^{(n)\frac{1}{2}}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}\right)\right)^2.
\end{aligned}\tag{29}$$

By Cauchy-Schwarz's inequality and $(a+b)^2\leq 2(a^2+b^2)$,

$$\begin{aligned}
&\left(\sum_{m\in I_0}\left(1+\frac{\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}}\right)\left(\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_3^{(n)\frac{1}{2}}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}\right)\right)^2\\
\leq&\sum_{m\in I_0}\left(1+\frac{\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}}\right)^2\sum_{m\in I_0}\left(\|f_m^*-\hat{f}_m\|_{L_2(\Pi)}+\lambda_3^{(n)\frac{1}{2}}\|f_m^*-\hat{f}_m\|_{\mathcal{H}_m}\right)^2
\end{aligned}$$

$$\leq 4 \sum_{m\in I_0} \left(1 + \frac{{\lambda_3^{(n)}}^{1+q} \|g_m^*\|_{\mathcal{H}_m}^2}{{\lambda_1^{(n)}}^2}\right) \sum_{m\in I_0} \left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right).$$

Thus the RHS of (29) is further bounded by

$$256 \max(\phi_s \sqrt{n}\xi_n^2, r) \left(d + \frac{{\lambda_3^{(n)}}^{1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{{\lambda_1^{(n)}}^2}\right) \sum_{m\in I_0} \left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right)$$

$$\overset{(4)}{\leq} 256 \max(\phi_s \sqrt{n}\xi_n^2, r) \left(d + \frac{{\lambda_3^{(n)}}^{1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{{\lambda_1^{(n)}}^2}\right) \left(\frac{\|\hat{f} - f^*\|_{L_2(\Pi)}^2}{(1-\rho(I_0)^2)\kappa(I_0)} + \sum_{m\in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right), \quad (30)$$

where we used Eq. (4) in Lemma 1 in the last line. By the assumption (25), we have $128 \max(\phi_s \sqrt{n}\xi_n^2, r) \left(d + \frac{{\lambda_3^{(n)}}^{1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{{\lambda_1^{(n)}}^2}\right) / (1-\rho(I_0)^2)\kappa(I_0) \leq \frac{1}{8}$. Hence the RHS of the above inequality is bounded by $\frac{1}{8}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right)$.

Step 2. (Bound of the second term (ii) in the RHS of Eq. (27)) By Eq. (28), we have on the event $\mathscr{E}_1$

$$\begin{aligned}
&\sum_{m=1}^M \eta(t)\phi_s\xi_n \left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}\right) \\
&\leq \sum_{m\in I_0} 8 \left(1 + \frac{{\lambda_3^{(n)}}^{\frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m}}{\lambda_1^{(n)}}\right) \eta(t)\phi_s\xi_n \left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + {\lambda_3^{(n)}}^{\frac{1}{2}} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}\right) \\
&\leq \frac{256\phi_s^2\eta(t)^2\xi_n^2}{(1-\rho(I_0)^2)\kappa(I_0)} \left(d + \frac{{\lambda_3^{(n)}}^{1+q}}{{\lambda_1^{(n)}}^2} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2\right) \\
&\quad + \frac{(1-\rho(I_0)^2)\kappa(I_0)}{16} \sum_{m\in I_0} \left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + {\lambda_3^{(n)}}^{\frac{1}{2}} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}\right)^2 \\
&\leq \frac{16}{(1-\rho(I_0)^2)\kappa(I_0)} \left(d{\lambda_1^{(n)}}^2 + {\lambda_3^{(n)}}^{1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2\right) \\
&\quad + \frac{1}{8}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right), \qquad (31)
\end{aligned}$$

where we used $(a+b)^2 \leq 2(a^2+b^2)$ and $(1-\rho(I_0)^2)\kappa(I_0) \sum_{m\in I_0} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 \leq \|\hat{f} - f^*\|_{L_2(\Pi)}^2$ (Lemma 1) in the last inequality.

Step 3. (Bound of the third term (iii) in the RHS of Eq. (27)) By Cauchy-Schwarz inequality and ${\lambda_3^{(n)}}^{\frac{1}{2}} = \lambda_2^{(n)}/\lambda_1^{(n)}$, we have

$$\begin{aligned}
&\sum_{m\in I_0} \frac{3}{2}(\lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}) \\
&\leq \frac{72}{2(1-\rho(I_0)^2)\kappa(I_0)} d{\lambda_1^{(n)}}^2 + \frac{(1-\rho(I_0)^2)\kappa(I_0)}{8} \sum_{m\in I_0} \left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right) \\
&\overset{(4)}{\leq} \frac{36}{(1-\rho(I_0)^2)\kappa(I_0)} d{\lambda_1^{(n)}}^2 + \frac{1}{8}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0} \lambda_3^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right). \qquad (32)
\end{aligned}$$

Step 4. (Bound of the last term (iv) in the RHS of Eq. (27)) By Eq. (61) in the proof of Lemma 9 (Appendix

G), and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\sum_{m\in I_0} 2\lambda_3^{(n)}\langle f_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} \overset{(61)}{\leq} 2\sum_{m\in I_0} {\lambda_3^{(n)}}^{\frac{1+q}{2}} \|g_m^*\|_{\mathcal{H}_m}(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + {\lambda_3^{(n)}}^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}) \\
\leq& \frac{16\sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{(1-\rho(I_0)^2)\kappa(I_0)}{\lambda_3^{(n)}}^{1+q} + \frac{(1-\rho(I_0)^2)\kappa(I_0)}{8}\sum_{m\in I_0}\left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + \lambda_3^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right) \\
\overset{(4)}{\leq}& \frac{16\sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2}{(1-\rho(I_0)^2)\kappa(I_0)}{\lambda_3^{(n)}}^{1+q} + \frac{1}{8}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0}\lambda_3^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right), \qquad (33)
\end{aligned}$$

where we used $\sum_{m\in I_0}\|g_m^*\|_{\mathcal{H}_m}^2 = \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2$ because $g_m^* = 0$ for $m \in I_0^c$.

Step 5. (Combining all the bounds) Substituting the inequalities (30), (31), (32) and (33) to Eq. (27), we obtain

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0}\lambda_3^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2 \\
\leq& \frac{16}{(1-\rho(I_0)^2)\kappa(I_0)}\left(d{\lambda_1^{(n)}}^2 + {\lambda_3^{(n)}}^{1+q}R_{g^*}^2\right) + \frac{36}{(1-\rho(I_0)^2)\kappa(I_0)}d{\lambda_1^{(n)}}^2 \\
&+ \frac{16R_{g^*}^2}{(1-\rho(I_0)^2)\kappa(I_0)}{\lambda_3^{(n)}}^{1+q} + \frac{1}{2}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0}\lambda_3^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right) \\
\leq& \frac{52}{(1-\rho(I_0)^2)\kappa(I_0)}\left(d{\lambda_1^{(n)}}^2 + {\lambda_3^{(n)}}^{1+q}\sum_{m=1}^M\|g_m^*\|_{\mathcal{H}_m}^2\right) \\
&+ \frac{1}{2}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0}\lambda_3^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right).
\end{aligned}$$

Moving the term $\frac{1}{2}\left(\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \sum_{m\in I_0}\lambda_3^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2\right)$ in the RHS to the left, we obtain the assertion. The probability bound is given by Theorems 7 and 8. □

Substituting $\lambda = d^{\frac{1}{1+q+s}}n^{-\frac{1}{1+q+s}}R_{2,g^*}^{-\frac{2}{1+q+s}}$ into the bound in Theorem 10, we obtain the convergence rate of elastic-net MKL (7) in Theorem 2 as in the following Corollary 11.

Corollary 11. *Suppose Assumptions 1–5 are satisfied, and set*

$$\lambda = d^{\frac{1}{1+q+s}}n^{-\frac{1}{1+q+s}}R_{2,g^*}^{-\frac{2}{1+q+s}}.$$

Then there exist constants $\tilde{C}_1$, $\tilde{C}_2$ and ψ_s depending on $s, c, L, C_1, \rho(I_0), \kappa(I_0)$ such that if $\lambda_1^{(n)}$, $\lambda_2^{(n)}$ and $\lambda_3^{(n)}$ are set as $\lambda_1^{(n)} = \psi_s\eta(t)\xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda_1^{(n)}\lambda^{\frac{1}{2}}$, $\lambda_3^{(n)} = \lambda$, then for all n satisfying $\frac{\log(M)}{\sqrt{n}} \leq 1$ and

$$\tilde{C}_1\phi_s\sqrt{n}\xi_n(\lambda)^2 d \leq 1, \qquad (34)$$

we have

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\
\leq& \tilde{C}_2\left(d^{\frac{1+q}{1+q+s}}n^{-\frac{1+q}{1+q+s}}R_{2,g^*}^{\frac{2s}{1+q+s}} + \frac{d\log(M)}{n} + d^{\frac{q+s}{1+q+s}}n^{-\frac{1+q}{1+q+s}-\frac{q(1-s)}{(1+s)(1+q+s)}}R_{2,g^*}^{\frac{2}{1+q+s}}\right)\eta(t)^2. \qquad (35)
\end{aligned}$$

with probability $1 - \exp(-t) - \exp(-\zeta_n(\frac{1}{\tilde{C}_1 d}, \lambda))$ for all $t \geq 1$.

Proof. **(Corollary 11)** We set $\psi_s = 4\phi_s$ and suppose the following relation is met:

$$\frac{256\max(\phi_s\sqrt{n}\xi_n^2, r)\left(d + \frac{{\lambda_3^{(n)}}^{1+q}}{{\lambda_1^{(n)}}^2}\sum_{m=1}^M\|g_m^*\|_{\mathcal{H}_m}^2\right)}{(1-\rho(I_0)^2)\kappa(I_0)} \leq \frac{1}{8}. \qquad (36)$$

Then the assumptions for Theorem 10 are met. Once we assume the above condition (36) is satisfied (later we show this is satisfied), we can apply Theorem 10 that says

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \lesssim d{\lambda_1^{(n)}}^2 + {\lambda_3^{(n)}}^{1+q} R_{2,g^*}^2.$$

When $\lambda = d^{\frac{1}{1+q+s}} n^{-\frac{1}{1+q+s}} R_{2,g^*}^{-\frac{2}{1+q+s}}$,

$$\lambda_1^{(n)} = \psi_s \xi_n(\lambda)\eta(t) = \psi_s \left(\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}} \vee \sqrt{\frac{\log(M)}{n}} \right) \eta(t).$$

Therefore

$$\begin{aligned}
d{\lambda_1^{(n)}}^2 &= \psi_s^2 \left(\frac{d\lambda^{-s}}{n} \vee \frac{d\lambda^{-1}}{n^{\frac{2}{1+s}}} \vee \frac{d\log(M)}{n} \right) \eta(t)^2 \\
&= \psi_s^2 \left(\frac{d^{1-\frac{s}{1+q+s}} n^{\frac{s}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}}}{n} \vee \frac{d^{1-\frac{1}{1+q+s}} n^{\frac{1}{1+q+s}} R_{2,g^*}^{-\frac{2}{1+q+s}}}{n^{\frac{2}{1+s}}} \vee \frac{d\log(M)}{n} \right) \eta(t)^2 \\
&= \psi_s^2 \left(d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} \vee d^{\frac{q+s}{1+q+s}} n^{-\frac{1+q}{1+q+s} - \frac{q(1-s)}{(1+s)(1+q+s)}} R_{2,g^*}^{\frac{2}{1+q+s}} \vee \frac{d\log(M)}{n} \right) \eta(t)^2,
\end{aligned}$$

and

$${\lambda_3^{(n)}}^{1+q} R_{2,g^*}^2 = d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{-\frac{2(1+q)}{1+q+s}+2} = d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}}.$$

By Eq. (26) in Theorem 10, we have

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\
\leq &\frac{104}{(1-\rho(I_0))^2\kappa(I_0)} \left(d{\lambda_1^{(n)}}^2 + {\lambda_3^{(n)}}^{1+q} \sum_{m=1}^M \|g_m^*\|_{\mathcal{H}_m}^2 \right) \\
\leq &\frac{104(\psi_s^2+1)\eta(t)^2}{(1-\rho(I_0))^2\kappa(I_0)} \left(d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} + d^{\frac{q+s}{1+q+s}} n^{-\frac{1+q}{1+q+s} - \frac{q(1-s)}{(1+s)(1+q+s)}} R_{2,g^*}^{\frac{2}{1+q+s}} + \frac{d\log(M)}{n} \right).
\end{aligned}$$

Thus by setting $\tilde{C}_2$ as

$$\tilde{C}_2 = \frac{104(\psi_s^2+1)}{(1-\rho(I_0))^2\kappa(I_0)},$$

we obtain the inequality (35).

Finally we show the condition (34) yields the condition (36) for appropriately chosen $\tilde{C}_1$ and r. Note that

$$\frac{{\lambda_3^{(n)}}^{1+q}}{{\lambda_1^{(n)}}^2} R_{2,g^*}^2 \leq d^{\frac{1+q}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} \Big/ \psi_s^2 \left(d^{-\frac{s}{1+q+s}} n^{-\frac{1+q}{1+q+s}} R_{2,g^*}^{\frac{2s}{1+q+s}} \right) \eta(t)^2 \leq \frac{d}{\psi_s^2 \eta(t)^2}.$$

Here by the definitions (21) and (22) of $\eta(t)$ and ϕ_s, we have $\eta(t) \geq 1$ and $\psi_s = 4\phi_s \geq 1$. Thus

$$\frac{{\lambda_3^{(n)}}^{1+q}}{{\lambda_1^{(n)}}^2} R_{2,g^*}^2 \leq \frac{d}{\psi_s^2 \eta(t)^2} \leq d.$$

Therefore the condition (36) is satisfied if the following inequality holds:

$$\frac{256 \max(\phi_s \sqrt{n} \xi_n^2, r)\,(d+d)}{(1-\rho(I_0)^2)\kappa(I_0)} \leq \frac{1}{8}.$$

Thus by setting $\tilde{C}_1 = \frac{8 \times 512}{(1-\rho(I_0)^2)\kappa(I_0)}$ and $r = \frac{1}{\tilde{C}_1 d}$, the condition (34) gives the condition (36). Substituting this r into the claim of Theorem 10, we obtain the assertion. □

The next theorem and Corollary 13 give the convergence rate of L_1-MKL. Note that for the convergence rate of L_1-MKL, we don't include the Convolution Assumption (Assumption 2), that is, the smoothness parameter q does not appear in the rate of L_1-MKL.

Theorem 12. *Suppose Assumptions 1 and 3–5 are satisfied, Let* $\lambda_1^{(n)} = 4\phi_s\eta(t)\xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda^{\frac{1}{2}}\lambda_1^{(n)}$ *for arbitrary* $\lambda > 0$ *and* $\lambda_3^{(n)} = 0$. *In this setting, for all* n *and* r *satisfying* $\frac{\log(M)}{\sqrt{n}} \leq 1$ *and*

$$\frac{128\max(\phi_s\sqrt{n}\xi_n^2, r)d}{(1-\rho(I_0)^2)\kappa(I_0)} \leq \frac{1}{8}, \tag{37}$$

we have

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{606}{(1-\rho(I_0))^2\kappa(I_0)}\left[d{\lambda_1^{(n)}}^2 + \lambda\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2\right], \tag{38}$$

with probability $1-\exp(-t)-\exp(-\zeta_n(r,\lambda))$ *for all* $t \geq 1$.

Proof. **(Theorem 12)** Notice again that the assumption (37) implies $r \leq \frac{1}{8}$. Thus the assertion of Lemma 9 holds. We assume the events $\mathscr{E}_1(t)$ and $\mathscr{E}_2(r)$ are met. By Theorems 7 and 8, the probability of $\mathscr{E}_1(t)\cup\mathscr{E}_2(r)$ is bounded from below by $1-\exp(-t)-\exp(-\zeta_n(r,\lambda))$.

We start from Eq. (27) in the proof of Theorem 10 with $\lambda_3^{(n)} = 0$:

$$\begin{aligned}
&\|\hat{f} - f^*\|_{L_2(\Pi)}^2 + \frac{1}{2}\sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}) \\
&\leq \underbrace{\max(\phi_s\sqrt{n}\xi_n^2, r)\left(\sum_{m=1}^M(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})\right)^2}_{\text{(i)}} \\
&\quad + \underbrace{\sum_{m=1}^M \eta(t)\phi_s\xi_n(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})}_{\text{(ii)}} \\
&\quad + \sum_{m\in I_0}\underbrace{\frac{3}{2}(\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m})}_{\text{(iii)}}.
\end{aligned} \tag{39}$$

As in the proof of Theorem 10, we bound each term (i) to (iv) as follows.

Step 1. (Bound of the first term (i) in the RHS of Eq. (39)) By Eq. (24) in Lemma 9 and the relation $(a+b)^2 \leq 2(a^2+b^2)$, the term (i) on the RHS of Eq. (39) can be upper bounded as

$$\begin{aligned}
&\max(\phi_s\sqrt{n}\xi_n^2, r)\left(\sum_{m=1}^M(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})\right)^2 \\
\overset{(24)}{\leq}& \max(\phi_s\sqrt{n}\xi_n^2, r)\left(8\sum_{m\in I_0}\left(\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m^*\|_{\mathcal{H}_m}\right)\right)^2 \\
\leq& 128\max(\phi_s\sqrt{n}\xi_n^2, r)\left(\sum_{m\in I_0}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}\right)^2 + 128\max(\phi_s\sqrt{n}\xi_n^2, r)\lambda\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2 \\
\leq& \frac{128\max(\phi_s\sqrt{n}\xi_n^2, r)}{(1-\rho(I_0)^2)\kappa(I_0)}(1-\rho(I_0)^2)\kappa(I_0)d\sum_{m\in I_0}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2
\end{aligned}$$

$$+ 128\max(\phi_s\sqrt{n}\xi_n^2, r)\lambda\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2$$
$$\overset{(4)}{\leq}\frac{128\max(\phi_s\sqrt{n}\xi_n^2, r)d}{(1-\rho(I_0)^2)\kappa(I_0)}\|f^*-\hat{f}\|_{L_2(\Pi)}^2 + 128\max(\phi_s\sqrt{n}\xi_n^2, r)d\frac{\lambda}{d}\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2, \tag{40}$$

where we used $(1-\rho(I_0)^2)\kappa(I_0)\sum_{m\in I_0}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 \leq \|\hat{f}-f^*\|_{L_2(\Pi)}^2$ (Eq. (4) in Lemma 1) in the last inequality. By the assumption (37), we have $\frac{128\max(\phi_s\sqrt{n}\xi_n^2, r)d}{(1-\rho(I_0)^2)\kappa(I_0)} \leq \frac{1}{8}$. Hence the RHS of the above inequality is bounded by $\frac{1}{8}\left[\|\hat{f}-f^*\|_{L_2(\Pi)}^2 + \frac{\lambda}{d}(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m})^2\right]$.

Step 2. (Bound of the second term (ii) in the RHS of Eq. (39)) By Eq. (24) in Lemma 9 and the relations $\eta(t)\phi_s\xi_n = \lambda_1^{(n)}/4$, $\lambda_2^{(n)} = \lambda^{\frac{1}{2}}\lambda_1^{(n)}$, we have on the event $\mathscr{E}_1$

$$\sum_{m=1}^M \eta(t)\phi_s\xi_n\left(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}\right)$$
$$= \sum_{m=1}^M \frac{1}{4}\left(\lambda_1^{(n)}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}\right)$$
$$\overset{(24)}{\leq} \sum_{m\in I_0} 2\left(\lambda_1^{(n)}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m}\right)$$
$$\leq \frac{8d{\lambda_1^{(n)}}^2}{(1-\rho(I_0)^2)\kappa(I_0)} + \frac{(1-\rho(I_0)^2)\kappa(I_0)}{8}\sum_{m\in I_0}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + 2\sum_{m\in I_0}\lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m}$$
$$\overset{(4)}{\leq} \frac{8d{\lambda_1^{(n)}}^2}{(1-\rho(I_0)^2)\kappa(I_0)} + \frac{1}{8}\|\hat{f}-f^*\|_{L_2(\Pi)}^2 + 2\sum_{m\in I_0}\lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m} \tag{41}$$

where we used $2ab \leq a^2+b^2$ in the third inequality and $(1-\rho(I_0)^2)\kappa(I_0)\sum_{m\in I_0}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 \leq \|\hat{f}-f^*\|_{L_2(\Pi)}^2$ (Eq. (4) in Lemma 1) in the last inequality.

Step 3. (Bound of the third term (iii) in the RHS of Eq. (39)) By Eq. (24), we have

$$\sum_{m\in I_0}\frac{3}{2}(\lambda_1^{(n)}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})$$
$$\leq \sum_{m=1}^M \frac{3}{2}(\lambda_1^{(n)}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})$$
$$\overset{(24)}{\leq} \sum_{m\in I_0} 12(\lambda_1^{(n)}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m})$$
$$\leq \frac{8\times 36d{\lambda_1^{(n)}}^2}{(1-\rho(I_0)^2)\kappa(I_0)} + \frac{(1-\rho(I_0)^2)\kappa(I_0)}{8}\sum_{m\in I_0}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 + 12\sum_{m\in I_0}\lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m}$$
$$\leq \frac{288d{\lambda_1^{(n)}}^2}{(1-\rho(I_0)^2)\kappa(I_0)} + \frac{1}{8}\|\hat{f}-f^*\|_{L_2(\Pi)}^2 + 12\sum_{m\in I_0}\lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m}, \tag{42}$$

where we used the same technique as in Eq. (41), that is, we used $2ab \leq a^2+b^2$ in the third inequality and $(1-\rho(I_0)^2)\kappa(I_0)\sum_{m\in I_0}\|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 \leq \|\hat{f}-f^*\|_{L_2(\Pi)}^2$ (Lemma 1) in the last inequality.

Step 4. (Combining all the bounds) Substituting the inequalities (40), (41) and (42) to Eq. (39), we obtain

$$\|\hat{f}-f^*\|_{L_2(\Pi)}^2$$

$$\leq \frac{8+288}{(1-\rho(I_0)^2)\kappa(I_0)} d{\lambda_1^{(n)}}^2 + (2+12)\lambda_2^{(n)} \sum_{m\in I_0} \|f_m^*\|_{\mathcal{H}_m} + \frac{\lambda}{8d}\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2 + \frac{3}{8}\|\hat{f}-f^*\|_{L_2(\Pi)}^2.$$

Moving the term $\frac{3}{8}\|\hat{f}-f^*\|_{L_2(\Pi)}^2$ in the RHS to the left, we obtain

$$\frac{1}{2}\|\hat{f}-f^*\|_{L_2(\Pi)}^2 \leq \frac{296}{(1-\rho(I_0)^2)\kappa(I_0)} d{\lambda_1^{(n)}}^2 + 14\lambda_2^{(n)} \sum_{m\in I_0} \|f_m^*\|_{\mathcal{H}_m} + \frac{\lambda}{8d}\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2.$$

Finally, since the relation $\lambda_2^{(n)} = \lambda_1^{(n)}\lambda^{\frac{1}{2}}$ yields

$$\lambda_2^{(n)} \sum_{m\in I_0} \|f_m^*\|_{\mathcal{H}_m} = \lambda_1^{(n)}\lambda^{\frac{1}{2}} \sum_{m\in I_0} \|f_m^*\|_{\mathcal{H}_m} \leq \frac{1}{2}\left[d{\lambda_1^{(n)}}^2 + \frac{\lambda}{d}\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2\right],$$

the last inequality indicates

$$\begin{aligned}\frac{1}{2}\|\hat{f}-f^*\|_{L_2(\Pi)}^2 &\leq \frac{296+7}{(1-\rho(I_0)^2)\kappa(I_0)} d{\lambda_1^{(n)}}^2 + \left(7+\frac{1}{8}\right)\frac{\lambda}{d}\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2 \\ &\leq \frac{303}{(1-\rho(I_0)^2)\kappa(I_0)}\left[d{\lambda_1^{(n)}}^2 + \frac{\lambda}{d}\left(\sum_{m\in I_0}\|f_m^*\|_{\mathcal{H}_m}\right)^2\right],\end{aligned}$$

where we used $(1-\rho(I_0)^2)\kappa(I_0) \leq 1$. □

Substituting $\lambda = d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{-\frac{2}{1+s}}$ which minimizes the bound obtained in Theorem 12 into the bound of Theorem 12 we obtain the convergence rate (6) of L_1-MKL in Theorem 2 as in the following Corollary 13.

Corollary 13. *Suppose Assumptions 1 and 3–5 are satisfied, and set*

$$\lambda = d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{-\frac{2}{1+s}}.$$

Then there exist constants $\tilde{C}_1$, $\tilde{C}_2$ and ψ_s depending on $s, c, L, C_1, \rho(I_0), \kappa(I_0)$ such that if $\lambda_1^{(n)}$, $\lambda_2^{(n)}$ and $\lambda_3^{(n)}$ are set as $\lambda_1^{(n)} = \psi_s \eta(t)\xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda_1^{(n)}\lambda^{\frac{1}{2}}$, $\lambda_3^{(n)} = 0$, then for all n satisfying $\frac{\log(M)}{\sqrt{n}}$ and

$$\tilde{C}_1\phi_s\sqrt{n}\xi_n(\lambda)^2 d \leq 1, \tag{43}$$

we have

$$\|\hat{f}-f^*\|_{L_2(\Pi)}^2 \leq \tilde{C}_2\left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} + d^{\frac{s-1}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2}{1+s}} + \frac{d\log(M)}{n}\right)\eta(t)^2. \tag{44}$$

with probability $1-\exp(-t)-\exp(-\zeta_n(\frac{1}{\tilde{C}_1 d},\lambda))$ for all $t \geq 1$.

Proof. **(Corollary 13)** The proof is similar to that of Corollary 11. Suppose we set $\psi_s = 4\phi_s$ and

$$\frac{128\max(\phi_s\sqrt{n}\xi_n^2, r)d}{(1-\rho(I_0)^2)\kappa(I_0)} \leq \frac{1}{8} \tag{45}$$

is satisfied. This inequality (45) is met from Eq. (43) if we set $\tilde{C}_1 = \frac{8\times 128}{(1-\rho(I_0)^2)\kappa(I_0)}$ and $r = \frac{1}{\tilde{C}_1 d}$. Then the assumptions for Theorem 12 are satisfied. Therefore we can apply Theorem 12 that says the following inequality holds:

$$\|\hat{f}-f^*\|_{L_2(\Pi)}^2 \lesssim d{\lambda_1^{(n)}}^2 + \frac{\lambda}{d}R_{1,f^*}^2,$$

with probability $1-\exp(-t)-\exp(-\zeta_n(\frac{\tilde{C}_1}{d},\lambda))$ for all $t \geq 1$. When $\lambda = d^{\frac{2}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{-\frac{2}{1+s}}$,

$$\begin{aligned} d{\lambda_1^{(n)}}^2 &= \psi_s^2 \left(\frac{d\lambda^{-s}}{n} \vee \frac{d\lambda^{-1}}{n^{\frac{2}{1+s}}} \vee \frac{d\log(M)}{n} \right) \eta(t)^2 \\ &= \psi_s^2 \left(\frac{d^{1-\frac{2s}{1+s}} n^{\frac{s}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}}}{n} \vee \frac{d^{1-\frac{2}{1+s}} n^{\frac{1}{1+s}} R_{1,f^*}^{\frac{2}{1+s}}}{n^{\frac{2}{1+s}}} \vee \frac{d\log(M)}{n} \right) \eta(t)^2 \\ &= \psi_s^2 \left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} \vee d^{\frac{s-1}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2}{1+s}} \vee \frac{d\log(M)}{n} \right) \eta(t)^2, \end{aligned}$$

and

$$\frac{\lambda}{d} R_{1,f^*}^2 = d^{\frac{2}{1+s}-1} n^{-\frac{1}{1+s}} R_{1,f^*}^{2-\frac{2}{1+s}} = d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,g^*}^{\frac{2s}{1+s}}.$$

By Eq. (38) in Theorem 12, we have

$$\begin{aligned} &\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\ \leq &\frac{606}{(1-\rho(I_0))^2 \kappa(I_0)} \left[d{\lambda_1^{(n)}}^2 + \lambda \left(\sum_{m \in I_0} \|f_m^*\|_{\mathcal{H}_m} \right)^2 \right] \\ \leq &\frac{606(\psi_s^2+1)}{(1-\rho(I_0))^2 \kappa(I_0)} \left(d^{\frac{1-s}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2s}{1+s}} + d^{\frac{s-1}{1+s}} n^{-\frac{1}{1+s}} R_{1,f^*}^{\frac{2}{1+s}} + \frac{d\log(M)}{n} \right) \eta(t)^2. \end{aligned}$$

Thus by setting $\tilde{C}_2$ as

$$\tilde{C}_2 = \frac{606(\psi_s^2+1)}{(1-\rho(I_0))^2 \kappa(I_0)},$$

we obtain the assertion (44). □

Theorem 2 is immediately derived from the combination of Corollaries 11 and 13 by setting $\tilde{C}$ as the maximum of $\tilde{C}_2$ appeared in both corollaries.

F Proofs of Theorems 7 and 8

Here we give the proofs of Theorems 7 and 8. The proof shares the same spirit with Meier et al. (2009) and Koltchinskii and Yuan (2010), but we give the proofs for the sake of completeness.

For a Hilbert space $\mathcal{G} \subset L_2(P)$, let the *i-th entropy number* $e_i(\mathcal{G} \to L(P))$ be the infimum of $\epsilon > 0$ for which $\mathcal{N}(\epsilon, \mathcal{B}_\mathcal{G}, L_2(P)) \leq 2^{i-1}$, where $\mathcal{B}_\mathcal{G}$ is the unit ball of $\mathcal{G}$. One can check that if the spectral assumption (A3) holds, the i-th entropy number is bounded as

$$e_i(\mathcal{H}_m \to L_2(\Pi)) \leq \tilde{c} i^{-\frac{1}{2s}}. \tag{46}$$

where $\tilde{c}$ is a constant depends on s and c.

We denote by $\{\sigma_i\}_{i=1}^n$ the Rademacher random variable that is an i.i.d. random variable such that $\sigma_i \in \{\pm 1\}$. It is known that, for a set of measurable functions $\mathcal{F}$ that is separable with respect to ∞-norm, the *Rademacher complexity* $\mathrm{E}[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i)]$ of $\mathcal{F}$ bounds the supremum of the discrepancy between the empirical and population means of all functions $f \in \mathcal{F}$ (see Lemma 2.3.1 of van der Vaart and Wellner (1996)):

$$\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(x_i) - \mathrm{E}[f]) \right| \right] \leq 2\mathrm{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| \right], \tag{47}$$

where the expectations are taken for both $\{x_i\}_{i=1}^n$ and $\{\sigma_i\}_{i=1}^n$.

The following proposition is the key in our analysis.

Proposition 14. *Let $\mathcal{B}_{\delta,a,b} \subset \mathcal{H}_m$ be a set such that $\mathcal{B}_{\delta,a,b} = \{f_m \in \mathcal{H}_m \mid \|f_m\|_{L_2(\Pi)} \leq \delta, \|f_m\|_{\mathcal{H}_m} \leq a, \|f_m\|_\infty \leq b\}$. Assume the Spectral Assumption (A3), then there exist constants $\tilde{c}_s, C'_s$ depending only s and c such that*

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

Proof. **(Proposition 14)** Let D_n be the empirical distribution: $D_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. To bound empirical processes, a bound of the entropy number with respect to the empirical L_2-norm is needed. Corollary 7.31 of Steinwart and Christmann (2008) gives the following upper bound: under the condition (46), there exists a constant $c_s > 0$ only depending on s such that

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

Finally this and Theorem 7.16 of Steinwart and Christmann (2008) gives the assertion. □

Using Proposition 14 and the *peeling device*, we obtain the following lemma (see also Meier et al. (2009)).

Lemma 15. *Under the Spectral Assumption (Assumption 3), there exists a constant C_s depending only on s and C such that for all $\lambda > 0$*

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m : \|f_m\|_{\mathcal{H}_m} \leq 1} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}}\right] \leq C_s \left(\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}}\right).$$

Proof. **(Lemma 15)** Let $\mathcal{H}_m(\delta) := \{f_m \in \mathcal{H}_m \mid \|f_m\|_{\mathcal{H}_m} \leq 1, \|f_m\|_{L_2(\Pi)} \leq \delta\}$ and $z = 2^{1/s} > 1$. Then by noticing $\|f_m\|_\infty \leq \|f_m\|_{\mathcal{H}_m}$, Proposition 14 gives

$$\begin{aligned}
&\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m : \|f_m\|_{\mathcal{H}_m} \leq 1} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}}\right] \\
\leq &\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m(\lambda^{1/2})} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}}\right] + \sum_{k=1}^\infty \mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m(z^k\lambda^{1/2}) \setminus \mathcal{H}_m(z^{k-1}\lambda^{1/2})} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}}\right] \\
\leq &C'_s \left(\frac{\lambda^{\frac{1-s}{2}}\tilde{c}_s^s}{\lambda^{\frac{1}{2}}\sqrt{n}} \vee \frac{\tilde{c}_s^{\frac{2s}{1+s}}}{n^{\frac{1}{1+s}}\lambda^{\frac{1}{2}}}\right) + \sum_{k=0}^\infty C'_s \left(\frac{z^{k(1-s)}\lambda^{\frac{1-s}{2}}\tilde{c}_s^s}{\sqrt{n}z^k\lambda^{\frac{1}{2}}} \vee \frac{\tilde{c}_s^{\frac{2s}{1+s}}}{n^{\frac{1}{1+s}}z^k\lambda^{\frac{1}{2}}}\right) \\
= &C'_s \left(\tilde{c}_s^s \sqrt{\frac{\lambda^{-s}}{n}} \vee \tilde{c}_s^{\frac{2s}{1+s}} \left(\frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}}\right)\right) + \sum_{k=0}^\infty C'_s \left(\tilde{c}_s^s z^{-sk}\sqrt{\frac{\lambda^{-s}}{n}} \vee \tilde{c}_s^{\frac{2s}{1+s}} z^{-k} \left(\frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}}\right)\right) \\
\leq &2C'_s \left(\frac{1}{1-z^{-s}}\tilde{c}_s^s\sqrt{\frac{\lambda^{-s}}{n}} + \frac{1}{1-z^{-1}}\tilde{c}_s^{\frac{2s}{1+s}}\left(\frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}}\right)\right) = 2C'_s \left(2\tilde{c}_s^s\sqrt{\frac{\lambda^{-s}}{n}} + \frac{2^{1/s}}{2^{1/s}-1}\tilde{c}_s^{\frac{2s}{1+s}}\left(\frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}}\right)\right) \\
\leq &2C'_s \left(2\tilde{c}_s^s + \frac{2^{1/s}}{2^{1/s}-1}\tilde{c}_s^{\frac{2s}{1+s}}\right)\left(\sqrt{\frac{\lambda^{-s}}{n}} \vee \left(\frac{\lambda^{-\frac{1}{2}}}{n^{\frac{1}{1+s}}}\right)\right).
\end{aligned}$$

By setting $C_s \leftarrow 2C'_s \left(2\tilde{c}_s^s + \frac{2^{1/s}}{2^{1/s}-1}\tilde{c}_s^{\frac{2s}{1+s}}\right)$, we obtain the assertion. □

The above lemma immediately gives the following corollary.

Corollary 16. *Under the Spectral Assumption (Assumption 3), for all $\lambda > 0$*

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right] \leq C_s \left(\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}}\right),$$

where C_s is the constant appeared in the statement of Lemma 15, and we employed a convention such that $\frac{0}{0} = 0$.

Proof. **(Corollary 16)** Dividing the denominator and the numerator in the supremand in the LHS by $\|f_m\|_{\mathcal{H}_m}$, the inequality reduces to Lemma 15. □

This corollary gives the following lemma.

Lemma 17. *Under the Spectral Assumption (Assumption 3), for all $\lambda > 0$*

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right] \leq 2C_s L \left(\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}}\right),$$

where C_s is the constant appeared in the statement of Lemma 15.

Proof. **(Lemma 17)** Here we write $Pf = \mathrm{E}[f]$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(x_i, y_i)$ for a function f. Notice that $P\epsilon f_m = 0$, thus $\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i) = (P_n - P)(\epsilon f_m)$. By contraction inequality (Ledoux and Talagrand, 1991, Theorem 4.12), we obtain

$$\begin{aligned}
&\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \frac{|(P - P_n)(\epsilon f_m)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right] \\
=&\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \left|(P - P_n)\frac{\epsilon f_m}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right|\right] \\
\overset{(47)}{\leq}& 2\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \left|\frac{\frac{1}{n}\sum_{n=1}^n \sigma_i \epsilon_i f_m(x_i)}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right|\right] \\
\leq& 2L\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \left|\frac{\frac{1}{n}\sum_{n=1}^n \sigma_i f_m(x_i)}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right|\right] \qquad (\because \text{contraction inequality}) \\
\leq& 2C_s L \left(\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda^{\frac{1}{2}} n^{\frac{1}{1+s}}}\right). \qquad (\because \text{Corollary 16})
\end{aligned}$$

This gives the assertion. □

From now on, we refer to C_s as the constant appeared in the statement of Lemma 15. We are ready to show Theorem 7 that gives the probability of $\mathscr{E}_1(t)$.

Proof. **(Theorem 7)** Since

$$\frac{\|f_m\|_{L_2(\Pi)}}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \leq 1, \tag{48}$$

$$\begin{aligned}
\frac{\|f_m\|_\infty}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} &\leq \frac{C_1\|f_m\|_{L_2(\Pi)}^{1-s}\|f_m\|_{\mathcal{H}_m}^s}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \overset{\text{Young}}{\leq} \frac{C_1\lambda^{-\frac{s}{2}}(\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \\
&\leq C_1\lambda^{-\frac{s}{2}},
\end{aligned} \tag{49}$$

applying Talagrand's concentration inequality (Proposition 6), we obtain

$$\begin{aligned}
&P\left(\sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \geq K\left[2C_s L\xi_n + \sqrt{\frac{L^2 t}{n}} + \frac{C_1 L\lambda^{-\frac{s}{2}} t}{n}\right]\right) \\
&\leq e^{-t}.
\end{aligned}$$

Therefore the uniform bound over all $m = 1, \ldots, M$ is given as

$$\begin{aligned}
&P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \geq K\left[2C_s L\xi_n + \sqrt{\frac{L^2 t}{n}} + \frac{C_1 L\lambda^{-\frac{s}{2}} t}{n}\right]\right) \\
&\leq \sum_{m=1}^M P\left(\sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \geq K\left[2C_s L\xi_n + \sqrt{\frac{L^2 t}{n}} + \frac{C_1 L\lambda^{-\frac{s}{2}} t}{n}\right]\right) \\
&\leq M e^{-t}.
\end{aligned}$$

Setting $t \leftarrow t + \log(M)$, we have

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \geq K\left[2C_s L\xi_n + \sqrt{\frac{L^2(t+\log(M))}{n}} + \frac{C_1 L\lambda^{-\frac{s}{2}}(t+\log(M))}{n}\right]\right) \leq e^{-t}. \tag{50}$$

Now

$$\begin{aligned}\sqrt{\frac{L^2(t+\log(M))}{n}} + \frac{C_1 L\lambda^{-\frac{s}{2}}(t+\log(M))}{n} &\leq L\sqrt{\frac{t}{n}} + L\sqrt{\frac{\log(M)}{n}} + \frac{C_1 L\lambda^{-\frac{s}{2}}}{\sqrt{n}}\left(\frac{t}{\sqrt{n}} + \frac{\log(M)}{\sqrt{n}}\right) \\ &\leq \xi_n\left(L\sqrt{t} + L + C_1 L\frac{t}{\sqrt{n}} + C_1 L\right) \leq \xi_n\left(2L + 2C_1 L\right)\eta(t).\end{aligned}$$

where we used $\frac{\log(M)}{\sqrt{n}} \leq 1$ in the second inequality. Thus Eq. (50) implies

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \geq K(2C_s L + 2L + 2C_1 L)\xi_n\eta(t)\right) \leq e^{-t}.$$

By substituting $\tilde{\phi}_s \leftarrow 2KL(C_s + 1 + C_1)$, we obtain

$$P\left(\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}} \geq \tilde{\phi}_s\xi_n\eta(t)\right) \leq e^{-t}. \tag{51}$$

Since $\tilde{\phi}_s \leq \phi_s$ by the definition, we obtain the assertion. □

By Theorem 7, we obtain the expectation of the quantity $\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}$, as in the following lemma.

Lemma 18. *Under the Basic Assumption, the Spectral Assumption and the Supnorm Assumption, when* $\frac{\log(M)}{\sqrt{n}} \leq 1$*, we have for all* $\lambda > 0$

$$\mathrm{E}\left[\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right] \leq 8K(C_s + 1 + C_1)\xi_n(\lambda). \tag{52}$$

Proof. Let $\tilde{\phi}_s = 2K(C_s + 1 + C_1)$. Substituting σ_i into ϵ_i in Eq. (51), Eq. (51) gives that

$$\begin{aligned}&\mathrm{E}\left[\max_m \sup_{f_m \in \mathcal{H}_m} \frac{|\frac{1}{n}\sum_{i=1}^n \sigma_i f_m(x_i)|}{\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \lambda\|f_m\|_{\mathcal{H}_m}^2}}\right] \leq \tilde{\phi}_s\xi_n + \sum_{t=0}^{\infty} e^{-t}\tilde{\phi}_s\xi_n\eta(t+1) \\ &\leq \tilde{\phi}_s\xi_n + \tilde{\phi}_s\xi_n\sum_{t=0}^{\infty} e^{-t}(t+1) \leq 4\tilde{\phi}_s\xi_n,\end{aligned}$$

where we used $\eta(t+1) = \max\{1, \sqrt{t+1}, (t+1)/\sqrt{n}\} \leq t+1$ in the second inequality. Thus we obtain the assertion. □

Next we show Theorem 8 that gives the probability bound of $\mathscr{E}_2(r)$.

Proof. **(Theorem 8)** By the symmetrization argument,

$$\mathrm{E}\left[\sup_{f_m \in \mathcal{H}_m} \frac{\left|\left\|\sum_{m=1}^M f_m\right\|_n^2 - \left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2\right|}{\left(\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}\right]$$

$$\overset{(47)}{\leq} 2\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(\sum_{m=1}^M f_m(x_i))^2\right|}{\left(\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}\right]$$
$$\leq \sup_{f_m\in\mathcal{H}_m}\frac{\left\|\sum_{m=1}^M f_m\right\|_\infty}{\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})}\times 2\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(\sum_{m=1}^M f_m(x_i))\right|}{\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})}\right], \tag{53}$$

where we used the contraction inequality in the last line (Ledoux and Talagrand, 1991, Theorem 4.12). Here we notice that

$$\begin{aligned}\left\|\sum_{m=1}^M f_m\right\|_\infty &\leq \sum_{m=1}^M C_1\|f_m\|_{L_2(\Pi)}^{1-s}\|f_m\|_{\mathcal{H}_m}^s \leq \sum_{m=1}^M C_1\lambda^{-\frac{s}{2}}\|f_m\|_{L_2(\Pi)}^{1-s}(\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}^2)^s\\ &\leq \sum_{m=1}^M C_1\lambda^{-\frac{s}{2}}[(1-s)\|f_m\|_{L_2(\Pi)}+s\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}]\\ &\leq \sum_{m=1}^M C_1\lambda^{-\frac{s}{2}}(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}),\end{aligned} \tag{54}$$

where we used Young's inequality $a^{1-s}b^s \leq (1-s)a+sb$ in the second line. Thus the RHS of the inequality (53) can be bounded as

$$\begin{aligned}&2C_1\lambda^{-\frac{s}{2}}\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\frac{\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(\sum_{m=1}^M f_m(x_i))\right|}{\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})}\right]\\ \leq&2C_1\lambda^{-\frac{s}{2}}\mathrm{E}\left[\sup_{f_m\in\mathcal{H}_m}\max_m\frac{\left|\frac{1}{n}\sum_{i=1}^n\sigma_i f_m(x_i)\right|}{\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}}\right],\end{aligned}$$

where we used the relation $\frac{\sum_m a_m}{\sum_m b_m}\leq\max_m(\frac{a_m}{b_m})$ for all $a_m\geq 0$ and $b_m\geq 0$ with a convention $\frac{0}{0}=0$. Therefore, by $\frac{\log(M)}{\sqrt{n}}\leq 1$ and Eq. (52), the right hand side is upper bounded by $8K(C_s+1+C_1)\lambda^{-\frac{s}{2}}\xi_n$. Here we again apply Talagrand's concentration inequality, then we have

$$\begin{aligned}&P\Bigg(\sup_{f_m\in\mathcal{H}_m}\frac{\left|\left\|\sum_{m=1}^M f_m\right\|_n^2-\left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2\right|}{\left(\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}\\ &\qquad\geq K\left[8K(C_s+1+C_1)\lambda^{-\frac{s}{2}}\xi_n+\sqrt{\frac{t}{n}}C_1\lambda^{-\frac{s}{2}}+\frac{C_1^2\lambda^{-s}t}{n}\right]\Bigg)\leq e^{-t},\end{aligned} \tag{55}$$

where we substituted the following upper bounds of B and U in Talagrand's inequality (17):

$$\begin{aligned}B=&\sup_{f_m\in\mathcal{H}_m}\mathrm{E}\left[\left(\frac{(\sum_{m=1}^M f_m)^2}{\left(\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}\right)^2\right]\\ \leq&\sup_{f_m\in\mathcal{H}_m}\mathrm{E}\left[\frac{(\sum_{m=1}^M f_m)^2}{\left(\sum_{m=1}^M\|f_m\|_{L_2(\Pi)}\right)^2}\frac{(\|\sum_{m=1}^M f_m\|_\infty)^2}{\left(\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}\right]\\ \overset{(54)}{\leq}&\sup_{f_m\in\mathcal{H}_m}\frac{\left(\sum_{m=1}^M\|f_m\|_{L_2(\Pi)}\right)^2}{\left(\sum_{m=1}^M\|f_m\|_{L_2(\Pi)}\right)^2}\frac{\left(\sum_{m=1}^M C_1\lambda^{-\frac{s}{2}}(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}{\left(\sum_{m=1}^M(\|f_m\|_{L_2(\Pi)}+\lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2}\\ \leq&C_1^2\lambda^{-s},\end{aligned}$$

where in the second inequality we used the relation $\mathrm{E}[(\sum_{m=1}^M f_m)^2] = \mathrm{E}[\sum_{m,m'=1}^M f_m f_{m'}] \leq \sum_{m,m'=1}^M \|f_m\|_{L_2(\Pi)}\|f_{m'}\|_{L_2(\Pi)} = (\sum_{m=1}^M \|f_m\|_{L_2(\Pi)})^2$, and

$$\begin{aligned}
U =& \sup_{f_m \in \mathcal{H}_m} \left\| \frac{(\sum_{m=1}^M f_m)^2}{\left(\sum_{m=1}^M (\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2} \right\|_\infty \\
\overset{(54)}{\leq}& \sup_{f_m \in \mathcal{H}_m} \frac{(\sum_{m=1}^M C_1 \lambda^{-\frac{s}{2}}(\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m}))^2}{\left(\sum_{m=1}^M (\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2} \\
\leq& C_1^2 \lambda^{-s},
\end{aligned}$$

where we used Eq. (54) in the second line. Now notice that

$$\begin{aligned}
& K\left[2C_1(2C_s + (C_1+1)K)\lambda^{-\frac{s}{2}}\xi_n + \sqrt{\frac{t}{n}}C_1\lambda^{-\frac{s}{2}} + \frac{C_1^2\lambda^{-s}t}{n}\right] \\
\leq& \sqrt{n}K\left[2C_1(2C_s + (C_1+1)K)\frac{\lambda^{-\frac{s}{2}}}{\sqrt{n}}\xi_n + \sqrt{\frac{t}{\log(M)}}C_1\xi_n\sqrt{\frac{\log(M)}{n}} + \frac{C_1^2\xi_n^2 t}{\sqrt{n}}\right] \\
\leq& \sqrt{n}K\left[2C_1(2C_s + (C_1+1)K) + \sqrt{\frac{t}{\log(M)}}C_1 + \frac{C_1^2 t}{\sqrt{n}}\right]\xi_n^2.
\end{aligned}$$

Therefore Eq. (55) gives the following inequality

$$\begin{aligned}
& \sup_{f_m \in \mathcal{H}_m} \frac{\left| \left\|\sum_{m=1}^M f_m\right\|_n^2 - \left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2 \right|}{\left(\sum_{m=1}^M (\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2} \\
\leq& K\left[8K(C_s + 1 + C_1) + C_1 + C_1^2\right]\sqrt{n}\xi_n^2 \max(1, \sqrt{t/\log(M)}, t/\sqrt{n}).
\end{aligned}$$

with probability $1 - \exp(-t)$. Since $K\left[8K(C_s + 1 + C_1) + C_1 + C_1^2\right] \leq \phi_s$ from the definition of ϕ_s, by substituting $t = \zeta_n(r, \lambda)$ into this bound, we obtain

$$\begin{aligned}
& \sup_{f_m \in \mathcal{H}_m} \frac{\left| \left\|\sum_{m=1}^M f_m\right\|_n^2 - \left\|\sum_{m=1}^M f_m\right\|_{L_2(\Pi)}^2 \right|}{\left(\sum_{m=1}^M (\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})\right)^2} \\
\leq& \phi_s\sqrt{n}\xi_n^2 \max(1, r/\phi_s\sqrt{n}\xi_n^2) = \max(\phi_s\sqrt{n}\xi_n^2, r).
\end{aligned}$$

This gives the assertion. □

G Proof of Lemma 9

On the event $\mathscr{E}_2(r)$, for all $f_m \in \mathcal{H}_m$ we obtain the upper bound of the regularization term as

$$\begin{aligned}
& \lambda_1^{(n)}\|f_m\|_n + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
& \leq \lambda_1^{(n)}\sqrt{\|f_m\|_{L_2(\Pi)}^2 + \max(\phi_s\sqrt{n}\xi_n^2, r)(\|f_m\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|f_m\|_{\mathcal{H}_m})^2} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
& \leq \lambda_1^{(n)}\sqrt{\|f_m\|_{L_2(\Pi)}^2 + 2\max(\phi_s\sqrt{n}\xi_n^2, r)(\|f_m\|_{L_2(\Pi)}^2 + \lambda\|f_m\|_{\mathcal{H}_m}^2)} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
& \leq \lambda_1^{(n)}\sqrt{\frac{5}{4}\|f_m\|_{L_2(\Pi)}^2 + \frac{\lambda}{4}\|f_m\|_{\mathcal{H}_m}^2} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
& \leq \lambda_1^{(n)}\sqrt{\frac{5}{4}}\|f_m\|_{L_2(\Pi)} + \frac{\lambda_1^{(n)}\lambda^{\frac{1}{2}}}{2}\|f_m\|_{\mathcal{H}_m} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m}
\end{aligned} \tag{56}$$

$$\leq \frac{3}{2}\left(\lambda_1^{(n)}\|f_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m}\right), \tag{57}$$

because $2\max(\phi_s\sqrt{n}\xi_n^2, r) \leq \frac{1}{4}$ and $\lambda^{\frac{1}{2}}\lambda_1^{(n)} = \lambda_2^{(n)}$. On the other hand, we also obtain a lower bound as

$$\begin{aligned}
&\lambda_1^{(n)}\|f_m\|_n + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
&\geq \lambda_1^{(n)}\sqrt{\max\{\|f_m\|_{L_2(\Pi)}^2 - 2\max(\phi_s\sqrt{n}\xi_n^2, r)(\|f_m\|_{L_2(\Pi)}^2 + \lambda\|f_m\|_{\mathcal{H}_m}^2), 0\}} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
&\geq \lambda_1^{(n)}\sqrt{\max\left\{\frac{3}{4}\|f_m\|_{L_2(\Pi)}^2 - \frac{1}{4}\lambda\|f_m\|_{\mathcal{H}_m}^2, 0\right\}} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
&= \sqrt{\max\left\{\frac{3}{4}{\lambda_1^{(n)}}^2\|f_m\|_{L_2(\Pi)}^2 - \frac{1}{4}{\lambda_2^{(n)}}^2\|f_m\|_{\mathcal{H}_m}^2, 0\right\}} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m} \\
&\geq \frac{1}{2}\left(\lambda_1^{(n)}\|f_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m\|_{\mathcal{H}_m}\right),
\end{aligned} \tag{58}$$

where in the last inequality we used the relation $\sqrt{\max(\frac{3}{4}a^2 - \frac{1}{4}b^2, 0)} \geq \frac{a-b}{2}$ for all $a, b \geq 0$ (this is because, when $a \geq b$, we have $\sqrt{\max(\frac{3}{4}a^2 - \frac{1}{4}b^2, 0)} \geq \sqrt{\max(\frac{1}{4}a^2 + \frac{1}{4}b^2, 0)} \geq \frac{1}{2}(a-b)$).

Note that, since $\hat{f}$ minimizes the objective function,

$$\begin{aligned}
&\|\hat{f} - f^*\|_n^2 + \sum_{m=1}^M (\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}^2) \\
\leq &\frac{1}{n}\sum_{n=1}^n\sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) + \sum_{m\in I_0}(\lambda_1^{(n)}\|f_m^*\|_n + \lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|f_m^*\|_{\mathcal{H}_m}^2).
\end{aligned} \tag{59}$$

Applying the inequalities $\sum_{m\in I_0}(\lambda_1^{(n)}\|f_m^*\|_n + \lambda_2^{(n)}\|f_m^*\|_{\mathcal{H}_m}) - \sum_{m\in I_0}(\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}) \leq \sum_{m\in I_0}(\lambda_1^{(n)}\|\hat{f}_m - f_m^*\|_n + \lambda_2^{(n)}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})$ and $\|f_m^*\|_{\mathcal{H}_m}^2 - \|\hat{f}_m\|_{\mathcal{H}_m}^2 = 2\langle f_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} - \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2$, the above inequality (59) yields

$$\begin{aligned}
&\|\hat{f} - f^*\|_n^2 + \sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}) \\
\leq &\frac{1}{n}\sum_{n=1}^n\sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) \\
&+ \sum_{m\in I_0}[\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_n + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)}(2\langle f_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} - \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2)].
\end{aligned}$$

Thus on the event $\mathcal{E}_2(r)$, by Eq. (57) and Eq. (58), we have

$$\begin{aligned}
&\|\hat{f} - f^*\|_n^2 + \frac{1}{2}\sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}) \\
\leq &\frac{1}{n}\sum_{n=1}^n\sum_{m=1}^M \epsilon_i(\hat{f}_m(x_i) - f_m^*(x_i)) + \\
&\sum_{m\in I_0}\left[\frac{3}{2}(\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}) + \lambda_3^{(n)}(2\langle f_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} - \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2)\right].
\end{aligned}$$

Moreover on the event $\mathcal{E}_1(t)$, we have

$$\|\hat{f} - f^*\|_n^2 + \frac{1}{2}\sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m})$$

$$
\begin{aligned}
&\leq \sum_{m=1}^{M} \eta(t)\phi_s\xi_n(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}) + \\
&\quad \sum_{m\in I_0}\left[\frac{3}{2}(\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}) + \lambda_3^{(n)}(2\langle f_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} - \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}^2)\right] \qquad (60)\\
&\Rightarrow \\
&\quad \frac{1}{4}\sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}) \\
&\leq \sum_{m\in I_0}\left[\frac{7}{4}(\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}) + 2\lambda_3^{(n)}\langle T_m^{\frac{q}{2}} g_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m}\right],
\end{aligned}
$$

where we used the relation $\eta(t)\phi_s\xi_n = \lambda_1^{(n)}/4$ and $\lambda_2^{(n)} = \lambda_1^{(n)}\lambda^{\frac{1}{2}}$. Finally we bound the last term $\langle T_m^{\frac{q}{2}} g_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m}$. By the Young's inequality for positive symmetric operator, we have

$$
\begin{aligned}
\lambda_3^{(n)1-q} T_m^q &= \lambda_3^{(n)\frac{1}{2}}\left(\lambda_3^{(n)-\frac{1}{2}} T_m \lambda_3^{(n)-\frac{1}{2}}\right)^q \lambda_3^{(n)\frac{1}{2}} \\
&\preceq qT_m + (1-q)\lambda_3^{(n)}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\lambda_3^{(n)}\langle f_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} \\
=&\lambda_3^{(n)}\langle T_m^{\frac{q}{2}} g_m^*, f_m^* - \hat{f}_m\rangle_{\mathcal{H}_m} \\
\leq&\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}\|\lambda_3^{(n)\frac{1-q}{2}} T_m^{\frac{q}{2}}(f_m^* - \hat{f}_m)\|_{\mathcal{H}_m} \\
\leq&\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}\sqrt{\langle f_m^* - \hat{f}_m, \left(qT_m + (1-q)\lambda_3^{(n)}\right) f_m^* - \hat{f}_m\rangle} \\
=&\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}\sqrt{q\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + (1-q)\lambda_3^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} \\
\leq&\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}\sqrt{\|f_m^* - \hat{f}_m\|_{L_2(\Pi)}^2 + \lambda_3^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}^2} \\
\leq&\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}(\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_3^{(n)\frac{1}{2}}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}). \qquad (61)
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
&\frac{1}{4}\sum_{m\in I_0^c}(\lambda_1^{(n)}\|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}) \\
&\leq \sum_{m\in I_0}\Bigg[\frac{7}{4}(\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}) \\
&\qquad\qquad + 2\lambda_3^{(n)\frac{1+q}{2}}\|g_m^*\|_{\mathcal{H}_m}\left(\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_3^{(n)\frac{1}{2}}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}\right)\Bigg],
\end{aligned}
$$

with probability $1 - \exp(-t) - \exp(-\zeta_n(r,\lambda))$. Adding $\frac{1}{4}\sum_{m\in I_0}(\lambda_1^{(n)}\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)}\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m})$ to both LHS and RHS of this inequality, the first assertion (23) is obtained.

Next we show the second assertion (24). On the event $\mathcal{E}_1(t)$, (59) yields

$$
\begin{aligned}
&\|\hat{f} - f^*\|_n^2 + \sum_{m=1}^{M}(\lambda_1^{(n)}\|\hat{f}_m\|_n + \lambda_2^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)}\|\hat{f}_m\|_{\mathcal{H}_m}^2) \\
&\leq \sum_{m=1}^{M}\eta(t)\phi_s\xi_n(\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}}\|\hat{f}_m - f_m^*\|_{\mathcal{H}_m})
\end{aligned}
$$

$$+ \sum_{m \in I_0} (\lambda_1^{(n)} \|f_m^*\|_n + \lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2).$$

Applying $\|f_m^*\|_n - \|\hat{f}_m\|_n \leq \|\hat{f}_m - f_m^*\|_n$, this gives

$$\begin{aligned}
&\|\hat{f} - f^*\|_n^2 + \sum_{m \in I_0^c} \lambda_1^{(n)} \|\hat{f}_m\|_n + \sum_{m=1}^M (\lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2) \\
&\leq \sum_{m=1}^M \eta(t) \phi_s \xi_n (\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda^{\frac{1}{2}} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}) \\
&\quad + \sum_{m \in I_0} (\lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_n + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2).
\end{aligned}$$

Moreover, on the event $\mathcal{E}_2(r)$, applying Eq. (57), Eq. (58) and the relations $\lambda_1^{(n)}/4 = \eta(t)\phi_s\xi_n$ and $\lambda_2^{(n)} = \lambda_1^{(n)} \lambda^{\frac{1}{2}}$, we obtain

$$\begin{aligned}
&\sum_{m \in I_0^c} \frac{1}{2} (\lambda_1^{(n)} \|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}) + \sum_{m \in I_0} \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m} + \sum_{m=1}^M \lambda_3^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2 \\
&\leq \sum_{m=1}^M \frac{1}{4} (\lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m}) \\
&\quad + \sum_{m \in I_0} \left((\frac{3}{2} \lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \frac{1}{2} \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m}) + \lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right) \\
&= \sum_{m \in I_0^c} \frac{1}{4} (\lambda_1^{(n)} \|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}) + \sum_{m \in I_0} \left(\frac{7}{4} \lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \frac{3}{4} \lambda_2^{(n)} \|\hat{f}_m - f_m^*\|_{\mathcal{H}_m} \right) \\
&\quad + \sum_{m \in I_0} \left(\lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right) \\
&\leq \sum_{m \in I_0^c} \frac{1}{4} (\lambda_1^{(n)} \|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}) + \sum_{m \in I_0} \left(\frac{7}{4} \lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \frac{3}{4} \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m} \right) \\
&\quad + \sum_{m \in I_0} \left(\frac{7}{4} \lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right),
\end{aligned}$$

where we used $\|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} \leq \|f_m^*\|_{\mathcal{H}_m} + \|\hat{f}_m\|_{\mathcal{H}_m}$ in the last inequality. Moving the terms $\|\hat{f}_m\|_{L_2(\Pi)}$ and $\|\hat{f}_m\|_{\mathcal{H}_m}$ in the RHS to the LHS, we have

$$\begin{aligned}
&\sum_{m \in I_0^c} \frac{1}{4} (\lambda_1^{(n)} \|\hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}) + \sum_{m \in I_0} \frac{1}{4} \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m} + \sum_{m=1}^M \lambda_3^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m}^2 \\
&\leq \sum_{m \in I_0} \left(\frac{7}{4} \lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \frac{7}{4} \lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right).
\end{aligned}$$

Since $f_m^* = 0$ for $m \in I_0^c$, adding $\sum_{m \in I_0} \frac{1}{4} \lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)}$ to both terms, this inequality yields

$$\begin{aligned}
&\sum_{m=1}^M \frac{1}{4} \left(\lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|\hat{f}_m\|_{\mathcal{H}_m} \right) \\
&\leq \sum_{m \in I_0} \left(2\lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + \frac{7}{4} \lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right).
\end{aligned}$$

Finally by the relation $\|\hat{f}_m\|_{\mathcal{H}_m} \geq \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} - \|f_m^*\|_{\mathcal{H}_m}$, we obtain

$$\sum_{m=1}^M \frac{1}{4} \left(\lambda_1^{(n)} \|f_m^* - \hat{f}_m\|_{L_2(\Pi)} + \lambda_2^{(n)} \|f_m^* - \hat{f}_m\|_{\mathcal{H}_m} \right)$$

$$\leq \sum_{m \in I_0} \left(2\lambda_1^{(n)} \|\hat{f}_m - f_m^*\|_{L_2(\Pi)} + 2\lambda_2^{(n)} \|f_m^*\|_{\mathcal{H}_m} + \lambda_3^{(n)} \|f_m^*\|_{\mathcal{H}_m}^2 \right).$$

Thus we obtain the second assertion (24).

H Proof of Theorem 3

Proof. **(Theorem 3)** The δ-packing number $\mathcal{M}(\delta, \mathcal{G}, L_2(P))$ of a function class $\mathcal{G}$ with respect to $L_2(P)$ norm is the largest number of functions $\{f_1, \ldots, f_{\mathcal{M}}\} \subseteq \mathcal{G}$ such that $\|f_i - f_j\|_{L_2(P)} \geq \delta$ for all $i \neq j$. It is easily checked that

$$\mathcal{N}(\delta/2, \mathcal{G}, L_2(P)) \leq \mathcal{M}(\delta, \mathcal{G}, L_2(P)) \leq \mathcal{N}(\delta, \mathcal{G}, L_2(P)). \tag{62}$$

First we give the assertion about the ℓ_∞-mixed-norm ball ($p = \infty$). To simplify the notation, set $R = R_\infty$. For a given $\delta_n > 0$ and $\varepsilon_n > 0$, let Q be the δ_n packing number $\mathcal{M}(\delta_n, \mathcal{H}_{\ell_\infty}^{d,q}(R), L_2(\Pi))$ of $\mathcal{H}_{\ell_\infty}^{d,q}(R)$ and N be the ε_n covering number $\mathcal{N}(\varepsilon_n, \mathcal{H}_{\ell_\infty}^{d,q}(R), L_2(\Pi))$ of $\mathcal{H}_{\ell_\infty}^{d,q}(R)$. Raskutti et al. (2010) utilized the techniques developed by Yang and Barron (1999) to show the following inequality in their proof of Theorem 2(b) :

$$\begin{aligned}
\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R)} \mathrm{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] &\geq \inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R)} \frac{\delta_n^2}{2} P[\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \geq \delta_n^2/2] \\
&\geq \frac{\delta_n^2}{2} \left(1 - \frac{\log(N) + \frac{n}{2\sigma^2}\varepsilon_n^2 + \log(2)}{\log(Q)} \right).
\end{aligned}$$

Now let $\tilde{Q}_m := \mathcal{M}\left(\delta_n/\sqrt{d}, \mathcal{H}_m^q(R), L_2(\Pi)\right)$ (remind the definition of $\mathcal{H}_m^q(R)$ (Eq. (14)), and since now $\mathcal{H}_m$ is taken as $\tilde{\mathcal{H}}$ for all m, the value $\tilde{Q}_m$ is common for all m). Thus by taking δ_n and ε_n to satisfy

$$\frac{n}{2\sigma^2}\varepsilon_n^2 \leq \log(N), \tag{63}$$

$$4\log(N) \leq \log(Q), \tag{64}$$

the minimax rate is lower bounded by $\frac{\delta_n^2}{4}$. In Lemma 5 of Raskutti et al. (2010), it is shown that if $\tilde{Q}_1 \geq 2$ and $d \leq M/4$, we have

$$\log(Q) \sim d\log(\tilde{Q}_1) + d\log\left(\frac{M}{d}\right).$$

By the estimation of the covering number of $\mathcal{H}_m^q(1)$ (Eq. (15)), the strong spectrum assumption (Eq. (10)) and the relation (62), we have

$$\log(\tilde{Q}_1) \sim \left(\frac{\delta_n}{R\sqrt{d}}\right)^{-2\frac{s}{1+q}} = \left(\frac{\delta_n}{R\sqrt{d}}\right)^{-2\tilde{s}}.$$

Thus the conditions (64) and (63) are satisfied if we set $\delta_n = C\varepsilon_n$ with an appropriately chosen constant C and we take ε_n so that the following inequality holds:

$$n\varepsilon_n^2 \lesssim d^{1+\tilde{s}} R^{2\tilde{s}} \varepsilon_n^{-2\tilde{s}} + d\log\left(\frac{M}{d}\right).$$

It suffices to take

$$\varepsilon_n^2 \sim d n^{-\frac{1}{1+\tilde{s}}} R^{\frac{2\tilde{s}}{1+\tilde{s}}} + \frac{d\log\left(\frac{M}{d}\right)}{n}. \tag{65}$$

Note that we have taken $R \geq \sqrt{\frac{\log(M/d)}{n}}$, thus $\tilde{Q}_m \geq 2$ is satisfied if we take the constant in Eq. (65) appropriately. Thus we obtain the assertion for $p = \infty$.

Next we give the assertion about the ℓ_p-mixed-norm ball. To simplify the notation, set $R = R_p$. Since $\mathcal{H}_{\ell_p}^{d,q}(R) \supseteq \mathcal{H}_{\ell_\infty}^{d,q}(R/d^{\frac{1}{p}})$, we obtain

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_p}^{d,q}(R)} \mathrm{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] \geq \inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R/d^{\frac{1}{p}})} \mathrm{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2].$$

Here notice that we have $R/d^{\frac{1}{p}} \geq \sqrt{\frac{\log(M/d)}{n}}$ by assumption. Thus we can apply the assertion about the ℓ_∞-mixed-norm ball to bound the RHS of the just above display. We have shown that

$$\begin{aligned}\inf_{\hat{f}} \sup_{f^* \in \mathcal{H}_{\ell_\infty}^{d,q}(R/\sqrt{d})} \mathrm{E}[\|\hat{f} - f^*\|_{L_2(\Pi)}^2] &\gtrsim d n^{-\frac{1}{1+\tilde{s}}} (R/d^{\frac{1}{p}})^{\frac{2\tilde{s}}{1+\tilde{s}}} + \frac{d \log\left(\frac{M}{d}\right)}{n} \\ &= d^{1-\frac{2\tilde{s}}{p(1+\tilde{s})}} n^{-\frac{1}{1+\tilde{s}}} R^{\frac{2\tilde{s}}{1+\tilde{s}}} + \frac{d \log\left(\frac{M}{d}\right)}{n}.\end{aligned}$$

This gives the assertion. □