The Annals of Statistics
2013, Vol. 41, No. 3, 1381–1405
DOI: 10.1214/13-AOS1095

FAST LEARNING RATE OF MULTIPLE KERNEL LEARNING: TRADE-OFF BETWEEN SPARSITY AND SMOOTHNESS

BY TAIJI SUZUKI[1] AND MASASHI SUGIYAMA[2]

University of Tokyo and Tokyo Institute of Technology

We investigate the learning rate of multiple kernel learning (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 nonzero components of the ground truth is relatively small, and show sharper convergence rates than the learning rates have ever shown for both ℓ_1 and elastic-net regularizations. Our analysis reveals some relations between the choice of a regularization function and the performance. If the ground truth is smooth, we show a faster convergence rate for the elastic-net regularization with less conditions than ℓ_1-regularization; otherwise, a faster convergence rate for the ℓ_1-regularization is shown.

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, Smola and Williamson (2005), Argyriou et al. (2006), Bach (2009), Cortes, Mohri and Rostamizadeh (2009a), 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, Lanckriet and Jordan (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, for example,

Received December 2011; revised January 2013.

[1]Supported in part by MEXT KAKENHI 22700289, and the Aihara project, the FIRST program from JSPS, initiated by CSTP.

[2]Supported in part by the FIRST program.

MSC2010 subject classifications. Primary 62G08, 62F12; secondary 62J07.

Key words and phrases. Sparse learning, restricted isometry, elastic-net, multiple kernel learning, additive model, reproducing kernel Hilbert spaces, convergence rate, smoothness.

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 $d^{(1-s)/(1+s)}n^{-1/(1+s)} + d\log(M)/n$ under rather restrictive conditions, where n is the number of samples, d is the number of nonzero 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, van de Geer and Bühlmann (2009) gave a near optimal convergence rate $d(n/\log(M))^{-1/(1+s)}$. 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, van de Geer and Bühlmann (2009) and established the bound $dn^{-1/(1+s)} + d\log(M)/n$. 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, Mohri and Rostamizadeh (2009b), Kloft, Rückert and Bartlett (2010)]. Fast learning rate of MKL in nonsparse settings is given by Kloft and Blanchard (2012) for ℓ_p-mixed-norm regularization and by Suzuki (2011a, 2011b) for regularizations corresponding to arbitrary monotonically increasing norms.

In this paper, we focus on the sparse setting (i.e., the total number of kernels is large, but the number of nonzero 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,

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

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

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, Wainwright and Yu (2009, 2012)]. Our work follows this line and shows 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 nonzero 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, van de Geer and Bühlmann (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 investigate a relation between the sparsity and the smoothness. Roughly speaking, L_1-MKL generates a sparser solution while elastic-net MKL generates a smoother solution. When the smoothness q of the truth is small (say $q = 0$), we give a faster convergence rate of L_1-MKL than that of elastic-net MKL. On the other hand, if the truth is smooth, elastic-net MKL can make use of the smoothness of the truth. In that situation, 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.

TABLE 1
Relation between our analysis and existing analyses

	Penalty	**Smoothness (q)**	**Minimax optimality**	**Convergence rate**
KY (2008)	ℓ_1	$q = 1$	?	$d^{(1-s)/(1+s)} n^{-1/(1+s)} + \frac{d \log(M)}{n}$
MGB (2009)	elastic-net	$q = 0$	$\times$	$(\frac{\log(M)}{n})^{1/(1+s)} (d + R_{2,g^*}^2)$
KY (2010)	ℓ_1	$q = 0$	ℓ_∞-ball	$\frac{(d+R_{1,f^*})}{n^{1/(1+s)}} + \frac{d \log(M)}{n}$
This paper	elastic-net	$0 \le q \le 1$	ℓ_2-ball	$(\frac{d}{n})^{(1+q)/(1+q+s)} R_{2,g^*}^{2s/(1+q+s)} + \frac{d \log(M)}{n}$
	ℓ_1	$q = 0$	ℓ_1-ball	$\frac{d^{(1-s)/(1+s)}}{n^{1/(1+s)}} R_{1,f^*}^{2s/(1+s)} + \frac{d \log(M)}{n}$

2.1. *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 $\mathcal{H}_m (m = 1, \ldots, M)$ with a kernel k_m over $\mathcal{X} \times \mathcal{X}$.

The elastic-net MKL we consider in this paper is the version considered in Meier, van de Geer and Bühlmann (2009),

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

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[3] [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 $\mathbf{K}_m = (k_m(x_i, x_j))_{i,j}$, the regularizer in (1) is expressed as

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

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, Lanckriet and Jordan (2004), the coordinate descent algorithms [Meier, van de Geer and Bühlmann (2008)] or the alternating direction method of multipliers [Boyd et al. (2011)].

[3]There is another version of MKL with elastic-net 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 equation (1) because the above one is too loose to properly bound the irrelevant components of the estimated function.

2.2. *Notation and assumptions*. Here, we present several assumptions used in our theoretical analysis and prepare notation.

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 $\varepsilon_i := y_i - f^*(x_i)$ is bounded as $|\varepsilon_i| \le L$ (a.s.).

(A1-2) For each $m = 1, \ldots, M$, the kernel function k_m is continuous and $\sup_{X \in \mathcal{X}} |k_m(X, X)| \le 1$.

The first assumption in (A1-1) ensures the model $\mathcal{H}$ is correctly specified, and the technical assumption $|\varepsilon_i| < L$ allows $\varepsilon_i f$ to be Lipschitz continuous with respect to f. The assumption of correct specification can be relaxed to misspecified settings, and the bounded noise can be replaced with i.i.d. Gaussian noise as in Raskutti, Wainwright and Yu (2012). However, for the sake of simplicity, we assume these conditions. It is known that assumption (A1-2) gives the relation $\|f_m\|_\infty \le \|f_m\|_{\mathcal{H}_m}$; see Chapter 4 of Steinwart and Christmann (2008).

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

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

It is known that this operator is compact, positive and self-adjoint [see Theorem 4.27 of Steinwart and Christmann (2008)], and hence the spectral theorem shows that there exist an at most countable orthonormal system $\{\phi_{\ell,m}\}_{\ell=1}^\infty$ and eigenvalues $\{\mu_{\ell,m}\}_{\ell=1}^\infty$ such that

$$T_m f = \sum_{\ell=1}^\infty \mu_{\ell,m} \langle \phi_{\ell,m}, f \rangle_{L_2(\Pi)} \phi_{\ell,m} \tag{2}$$

for $f \in L_2(\Pi)$. Here we assume $\{\mu_{\ell,m}\}_{\ell=1}^\infty$ is sorted in descending order, that is, $\mu_{1,m} \ge \mu_{2,m} \ge \mu_{3,m} \ge \cdots \ge 0$. Associated with T_m, we can define an operator $\tilde{T}_m : \mathcal{H}_m \to \mathcal{H}_m$ as

$$\langle f_m', \tilde{T}_m f_m \rangle_{\mathcal{H}_m} = \mathrm{E}\big[f_m'(X) f_m(X)\big] = \left\langle f_m', \int k_m(\cdot, x) f_m(x) \, \mathrm{d}\Pi(x) \right\rangle_{\mathcal{H}_m}.$$

For the canonical inclusion map $\iota_m : \mathcal{H}_m \to L_2(\Pi)$, one can check that the following commutative relation holds:

$$T_m \iota_m f_m = \iota_m \tilde{T}_m f_m,$$

$$\begin{array}{ccc} \mathcal{H}_m & \xrightarrow{\tilde{T}_m} & \mathcal{H}_m \\ {\scriptstyle \iota_m}\downarrow & & \downarrow{\scriptstyle \iota_m} \\ L_2(\Pi) & \xrightarrow{T_m} & L_2(\Pi). \end{array}$$

Thus we use the same notation for T_m and $\tilde{T}_m$ and denote by T_m referring to both operators.

Due to Mercer's theorem [Ferreira and Menegatto (2009)], k_m has the following spectral expansion:

$$k_m(x, x') = \sum_{k=1}^{\infty} \mu_{k,m} \phi_{k,m}(x) \phi_{k,m}(x'),$$

where the convergence is absolute and uniform. Thus, the inner product of the RKHS $\mathcal{H}_m$ can be expressed as $\langle f_m, g_m \rangle_{\mathcal{H}_m} = \sum_{k=1}^{\infty} \mu_{k,m}^{-1} \langle f_m, \phi_{k,m} \rangle_{L_2(\Pi)} \times \langle \phi_{k,m}, g_m \rangle_{L_2(\Pi)}$.

The following assumption is regarding the smoothness of the true function f_m^*.

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

$$f_m^* = T_m^{q/2} g_m^*. \tag{A2}$$

We denote $(g_1^*, \ldots, g_M^*)$ and $\sum_{m=1}^{M} g_m^*$ by g^* (we use the same notation for both "vector" and "function" representations with a slight abuse of notation). The constant q represents the smoothness of the truth f_m^* because f_m^* is generated by operating the integral operator $T_m^{q/2}$ to g_m^* ($f_m^*(x) = \sum_{\ell=1}^{\infty} \mu_{\ell,m}^{q/2} \langle \phi_{\ell,m}, g_m^* \rangle_{L_2(\Pi)} \times \phi_{\ell,m}(x)$), and high-frequency components are suppressed as q becomes large. Therefore, as q becomes larger, f^* becomes "smoother." 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 assuming $q = 1$, and Bach (2008) showed the consistency of MKL under $q = 1$. 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, van de Geer and Bühlmann (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 that as q increases, the space of the functions that satisfy (A2) becomes "simpler." Thus, it might be natural to expect that, under 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. By Theorem 4.27 of Steinwart and Christmann (2008), the sum of $\mu_{\ell,m}$ is bounded ($\sum_\ell \mu_{\ell,m} < \infty$), and thus $\mu_{\ell,m}$ decreases with order ℓ^{-1} ($\mu_{\ell,m} = o(\ell^{-1})$). We further assume the sequence of the eigenvalues converges even faster to zero.

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

$$\mu_{j,m} \le c j^{-1/s}, \qquad (1 \le \forall j, 1 \le \forall m \le M), \tag{A3}$$

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

It was shown that spectral assumption (A3) gives a bound on the *entropy number* of the RKHSs [Steinwart, Hush and Scovel (2009)]. Remember that the ε-covering number $\mathcal{N}(\varepsilon, \mathcal{B}_\mathcal{G}, L_2(\Pi))$ with respect to $L_2(\Pi)$ for a Hilbert space $\mathcal{G}$ is the minimal number of balls with radius ε needed to cover the unit ball $\mathcal{B}_\mathcal{G}$ in $\mathcal{G}$ [van der Vaart and Wellner (1996)]. The ith entropy number $e_i(\mathcal{G} \to L_2(\Pi))$ is the infimum of $\varepsilon > 0$ for which $\mathcal{N}(\varepsilon, \mathcal{B}_\mathcal{G}, L_2(\Pi)) \le 2^{i-1}$. If spectral assumption (A3) holds, there exists a constant $\tilde{c}$ that depends only on s and c such that the ith entropy number is bounded as

$$e_i\bigl(\mathcal{H}_m \to L_2(\Pi)\bigr) \le \tilde{c} i^{-1/(2s)}, \tag{3}$$

and the converse is also true; see Theorem 15 of Steinwart, Hush and Scovel (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 entropy number in terms of the spectrum is provided in Appendix A. The entropy number of the space of functions that satisfy the Convolution assumption (A2) is also provided there.

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

ASSUMPTION 4 (Sup-norm assumption). Along with the spectral assumption (A3), there exists a constant C_1 such that

$$\|f_m\|_\infty \le C_1 \|f_m\|_{L_2(\Pi)}^{1-s} \|f_m\|_{\mathcal{H}_m}^{s} \qquad (\forall f_m \in \mathcal{H}_m, m = 1, \dots, M), \tag{A4}$$

where s is the exponent defined in 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 sup-norm assumption (A4). 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 (A4) 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, Hush and Scovel (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.

We denote by I_0 the indices of truly active kernels, that is,

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

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 $I \subseteq \{1, \dots, M\}$, we define $\mathcal{H}_I = \bigoplus_{m \in I} \mathcal{H}_m$, and denote by $f_I \in \mathcal{H}_I$ the restriction of f to an index set I, that is, $f_I = \sum_{m \in I} f_m$.

Now we introduce a geometric quantity that represents dependency between RKHSs. That quantity is related to the restricted eigenvalue condition [Bickel, Ritov and Tsybakov (2009)] and is required to show a nice convergence property of MKL. For a given set of indices $I \subseteq \{1, \dots, M\}$ and $b \geq 0$, we define

$$\beta_b(I) := \sup\Bigg\{\beta > 0 \Bigg| \beta \leq \frac{\|\sum_{m=1}^M f_m\|_{L_2(\Pi)}}{(\sum_{m \in I} \|f_m\|_{L_2(\Pi)}^2)^{1/2}}, \\ \forall f \in \mathcal{H} \text{ such that } b \sum_{m \in I} \|f_m\|_{L_2(\Pi)} \geq \sum_{m \notin I} \|f_m\|_{L_2(\Pi)}\Bigg\}.$$

For $I = I_0$, we abbreviate $\beta_b(I_0)$ as

$$\beta_b := \beta_b(I_0).$$

This quantity plays an important role in our analysis. Roughly speaking, this represents the correlation between RKHSs under the condition that the components within the relevant indices I well "dominate" the rest of the components. One can see that $\beta_b(I)$ is nonincreasing with respect to b. The quantity β_b is first introduced by Bickel, Ritov and Tsybakov (2009) to define the restricted eigenvalue condition in the context of parametric model such as the Lasso and the Dantzig selector. In the context of MKL, Koltchinskii and Yuan (2010) introduced this quantity to analyze a convergence rate of L_1-MKL. We will assume that $\beta_b(I_0)$ is bounded from below with some $b > 0$ so that 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.

Here we give a sufficient condition that $\beta_b(I)$ is bounded from below. 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,

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

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

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

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.

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

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

thus

$$\beta_\infty(I) \geq \sqrt{(1-\rho(I)^2)\kappa(I)}.$$

Koltchinskii and Yuan (2008) and Meier, van de Geer and Bühlmann (2009) analyzed statistical properties of MKL under the *incoherence condition* where $(1-\rho(I_0)^2)\kappa(I_0)$ is bounded from below, that is, RKHSs are not too dependent on each other. In this paper, we employ a less restrictive condition where β_b is bounded from below for some positive real b.

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 equation (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, Wainwright and Yu (2012)]. 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

$$\eta(t) := \max(1, \sqrt{t}, t/\sqrt{n}), \qquad \xi_n := \xi_n(\lambda) = \max\left(\frac{\lambda^{-s/2}}{\sqrt{n}}, \frac{\lambda^{-1/2}}{n^{1/(1+s)}}, \sqrt{\frac{\log(M)}{n}}\right).$$

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

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

Let

$$b_1 = 16\left(1 + \frac{\sqrt{d}\max_{m\in I_0}\|g_m^*\|_{\mathcal{H}_m}}{R_{2,g^*}}\right), \qquad b_2 = 16.$$

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 1–4 are satisfied. Then there exist constants $\tilde{C}_1, \tilde{C}_2$ and ψ_s depending on s, c, L, C_1 such that the following convergence rates hold*:

(Elastic-net MKL). *Set $\lambda_1^{(n)} = \psi_s \eta(t)\xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda_1^{(n)}\lambda^{1/2}$, $\lambda_3^{(n)} = \lambda$ where $\lambda = d^{1/(1+q+s)}n^{-1/(1+q+s)}R_{2,g^*}^{-2/(1+q+s)}$. Then for all n satisfying $\frac{\log(M)}{\sqrt{n}} \le 1$ and*

$$\frac{\tilde{C}_1}{\beta_{b_1}^2}\psi_s\sqrt{n}\xi_n(\lambda)^2 d \le 1, \tag{4}$$

the generalization error of elastic-net MKL is bounded as

$$\begin{aligned}\|\hat{f} - f^*\|_{L_2(\Pi)}^2 &\le \frac{\tilde{C}_2}{\beta_{b_1}^2}\Big(d^{(1+q)/(1+q+s)}n^{-(1+q)/(1+q+s)}R_{2,g^*}^{2s/(1+q+s)} \\ &\quad + d^{(q+s)/(1+q+s)}n^{-(1+q)/(1+q+s)-q(1-s)/((1+s)(1+q+s))} \\ &\qquad \times R_{2,g^*}^{2/(1+q+s)} + \frac{d\log(M)}{n}\Big)\eta(t)^2,\end{aligned} \tag{5}$$

with probability $1 - \exp(-t) - \exp(-\min\{\frac{\beta_{b_1}^4\log(M)}{\tilde{C}_1^2\psi_s^2 n\xi_n(\lambda)^4 d^2}, \frac{\beta_{b_1}^2}{\tilde{C}_1\psi_s\xi_n(\lambda)^2 d}\})$ for all $t \ge 1$.

(L_1-MKL). *Set $\lambda_1^{(n)} = \psi_s\eta(t)\xi_n(\lambda)$, $\lambda_2^{(n)} = \lambda_1^{(n)}\lambda^{1/2}$, $\lambda_3^{(n)} = 0$ where $\lambda = d^{(1-s)/(1+s)}n^{-1/(1+s)}R_{1,f^*}^{-2/(1+s)}$. Then for all n satisfying $\frac{\log(M)}{\sqrt{n}} \le 1$ and*

$$\frac{\tilde{C}_1}{\beta_{b_2}^2}\psi_s\sqrt{n}\xi_n(\lambda)^2 d \le 1, \tag{6}$$

the generalization error of L_1-MKL is bounded as

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \leq \frac{\tilde{C}_2}{\beta_{b_2}^2}\Big(d^{(1-s)/(1+s)}n^{-1/(1+s)}R_{1,f^*}^{2s/(1+s)} + d^{(s-1)/(1+s)}n^{-1/(1+s)}R_{1,f^*}^{2/(1+s)} + \frac{d\log(M)}{n}\Big)\eta(t)^2, \tag{7}$$

with probability $1 - \exp(-t) - \exp(-\min\{\frac{\beta_{b_2}^4\log(M)}{\tilde{C}_1^2\psi_s^2 n\xi_n(\lambda)^4 d^2}, \frac{\beta_{b_2}^2}{\tilde{C}_1\psi_s\xi_n(\lambda)^2 d}\})$ *for all* $t \geq 1$.

The proof of Theorem 2 is provided in Section S.3 of the supplementary material [Suzuki and Sugiyama (2013)]. The bounds presented in the theorem can be further simplified under additional conditions. To show simplified bounds, we assume that β_{b_1} and β_{b_2} are bounded from below by a positive constant; cf. the restricted eigenvalue condition, Bickel, Ritov and Tsybakov (2009). There exists $C_2 > 0$ such that $\beta_{b_2} \geq \beta_{b_1} \geq C_2$. This condition is satisfied if $\beta_{16(1+\sqrt{d})} \geq C_2$ because $\frac{\sqrt{d}\max_{m\in I_0}\|g_m^*\|_{\mathcal{H}_m}}{R_{2,g^*}} \leq \sqrt{d}$. Then we obtain simplified bounds with weak conditions. If $R_{1,f^*} \leq Cd$ with a constant C (this holds if $\|f_m^*\|_{\mathcal{H}_m} \leq C$ for all m), then the first term in the learning rate (7) of L_1-MKL dominates the second term, and thus equation (7) becomes

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

Similarly, as for the bound of elastic-net MKL, if $R_{2,g^*}^2 \leq Cn^{q/(1+s)}d$ with a constant C (this holds if $\|g_m^*\|_{\mathcal{H}_m} \leq \sqrt{C}$ for all m), then equation (5) becomes

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

Here notice that the tail probability can be bounded as

$$\exp\Big(-\min\Big\{\frac{\beta_{b_1}^4\log(M)}{\tilde{C}_1^2\psi_s^2 n\xi_n(\lambda)^4 d^2}, \frac{\beta_{b_1}^2}{\tilde{C}_1\psi_s\xi_n(\lambda)^2 d}\Big\}\Big) \leq \exp\big(-\min\{\log(M), \sqrt{n}\}\big) = \frac{1}{M},$$

under the conditions of equation (4) and $\frac{\log(M)}{\sqrt{n}} \leq 1$ [the same inequality also holds under equation (6), even if we replace β_{b_1} with β_{b_2}].

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 appear to be two convergence rates (8) and (9) that posses different characteristics depending on $\lambda_3^{(n)} = 0$, or not. There appears to be no dependency on the smoothness parameter q in bound (8) of L_1-MKL, while 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. The terms with respect to d are $d^{(1-s)/(1+s)}$ for L_1-MKL (8) and $d^{1/(1+s)}$ 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

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

This suggests that, when the truth is nonsmooth, 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 in the presented bound is $n^{-(1+q)/(1+q+s)}$ for elastic-net MKL that is faster than that of L_1-MKL ($n^{-1/(1+s)}$). We suggest that 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, we obtained a faster convergence bound of L_1-MKL than that of 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 bound of elastic-net MKL turns out to be faster than that of L_1-MKL [$d^{(1-s)/(1+s)} n^{-1/(1+s)} R_{1,f^*}^{2s/(1+s)} \geq d^{(1+q)/(1+q+s)} n^{-(1+q)/(1+q+s)} R_{2,g^*}^{2s/(1+q+s)}$]. This might indicate 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. Here, we would like to point out that the comparison between L_1-MKL and elastic-net MKL is just based on the upper bounds of the convergence rates. Thus there is still the possibility that L_1-MKL can also make use of the smoothness q of the true function to achieve a faster rate. We will give discussions about this issue in Section 6.

Finally, we give a comprehensive representation of Theorem 2 that gives a clear correspondence to the minimax optimal rate given in the next subsection.

COROLLARY 3. *Suppose the same condition as Theorem 2. Define* $\tilde{s} = \frac{s}{1+q}$. *Then there exists constant* $\tilde{C}'$ *depending on* s, c, L, C_1 *such that the following convergence rates hold*:

(Elastic-net MKL). *If* $1 \le R_{2,g^*}$ *and* $\|g_m^*\|_{\mathcal{H}_m} \le C$ $(\forall m \in I_0)$ *with a constant* C, *then for all* $p \ge 2$, *elastic-net MKL achieves the following convergence rate*:

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \le \frac{\tilde{C}'}{\beta_{b_1}^2}\left(d^{1-2\tilde{s}/(p(1+\tilde{s}))} n^{-1/(1+\tilde{s})} R_{p,g^*}^{2\tilde{s}/(1+\tilde{s})} + \frac{d\log(M)}{n}\right)\eta(t)^2,$$

with probability $1 - \exp(-t) - 1/M$ *for all* $t \ge 1$.

(L_1-MKL). *If* $1 \le R_{1,f^*}$ *and* $\|f_m^*\|_{\mathcal{H}_m} \le C$ $(\forall m \in I_0)$ *with a constant* C, *then for all* $p \ge 1$, L_1*-MKL achieves the following convergence rate*:

$$\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \le \frac{\tilde{C}'}{\beta_{b_2}^2}\left(d^{1-2s/(p(1+s))} n^{-1/(1+s)} R_{p,f^*}^{2s/(1+s)} + \frac{d\log(M)}{n}\right)\eta(t)^2,$$

with probability $1 - \exp(-t) - 1/M$ *for all* $t \ge 1$.

PROOF. Due to Jensen's inequality, we always have $R_{2,g^*} \le d^{1/2-1/p} R_{p,g^*}$ for $p \ge 2$ and $R_{1,f^*} \le d^{1-1/p} R_{p,f^*}$ for $p \ge 1$. Thus we have

$$d^{1/(1+\tilde{s})} n^{-1/(1+\tilde{s})} R_{2,g^*}^{2\tilde{s}/(1+\tilde{s})} \le d^{1-2\tilde{s}/(p(1+\tilde{s}))} n^{-1/(1+\tilde{s})} R_{p,g^*}^{2\tilde{s}/(1+\tilde{s})},$$

$$d^{(1-s)/(1+s)} n^{-1/(1+s)} R_{1,f^*}^{2s/(1+s)} \le d^{1-2s/(p(1+s))} n^{-1/(1+s)} R_{p,f^*}^{2s/(1+s)}.$$

Combining this and the discussions to derive equations (8) and (9), we have the assertion. □

Below, we show that 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 of* ℓ_p*-mixed-norm ball.* 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 induced from the same 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, that is, $\Pi = Q^M$. 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

$$c' j^{-1/s} \le \mu_j \le c j^{-1/s}, \tag{10}$$

where $\{\mu_j\}_j$ is the spectrum of the integral operator $T_{\tilde{k}}$ with respect to the kernel $\tilde{k}$; see equation (2). We also assume that the noise $\{\varepsilon_i\}_{i=1}^n$ 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 nonzero components in $\mathcal{H}$ defined by $\mathcal{H}_0(d) := \{(f_1, \ldots, f_M) \in \mathcal{H} \mid \#\{m \mid \|f_m\|_{\mathcal{H}_m} \neq 0\} \leq d\}$. We define the ℓ_p-mixed-norm ball ($p \geq 1$) with radius R in $\mathcal{H}_0(d)$ as

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

In Raskutti, Wainwright and Yu (2012), the minimax learning rate on $\mathcal{H}_{\ell_\infty}^{d,0}(R)$ (i.e., $p = \infty$ and $q = 0$) was derived.[4] 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 4. *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 $R_p \geq d^{1/p}\sqrt{\frac{\log(M/d)}{n}}$, there exists a constant $\widehat{C}_1$ such that*

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

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.

A proof of Theorem 4 is provided in Section S.7 of the supplementary material [Suzuki and Sugiyama (2013)].

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 (i.e., 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 $R_{p,f^*} := (\sum_m \|f_m^*\|_{\mathcal{H}_m}^p)^{1/p}$; then, as in the proof of Corollary 3, we always have

[4] The set $\mathcal{F}_{M,d,\mathcal{H}}(R)$ in Raskutti, Wainwright and Yu (2012) corresponds to $\mathcal{H}_{\ell_\infty}^{d,0}(R)$ in the current paper.

$R_{1,f^*} \le d^{1-1/p} R_{p,f^*} \le d R_{\infty,f^*}$ due to Jensen's inequality, and consequently we have

$$
\begin{aligned}
d^{(1-s)/(1+s)} n^{-1/(1+s)} R_{1,f^*}^{2s/(1+s)} &\le d^{1-2s/(p(1+s))} n^{-1/(1+s)} R_{p,f^*}^{2s/(1+s)} \\
&\le d n^{-1/(1+s)} R_{\infty,f^*}^{2s/(1+s)}.
\end{aligned} \tag{12}
$$

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 $0 \le q \le 1$, 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 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 $R_{p,f^*} = d^{1/p}$. Thus, both rates are $dn^{-1/(1+s)} + \frac{d\log(M)}{n}$; 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 we will see later, the existing bounds are the minimax rate on the ℓ_∞-mixed-norm ball and thus are tight only in the homogeneous setting.

4. 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* so that the estimator bounded in a way that makes the validation procedure effective. Given the estimator $\hat{f}_\Lambda$ and a positive real $B > 0$, the clipped estimator $\check{f}_\Lambda$ is given as

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

To appropriately choose B, we assume that we can roughly estimate the sup-norm $\|f^*\|_\infty$ of the true function, and B is set to satisfy $\|f^*\|_\infty < B$. 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 $\|f^*\|_\infty < B$. It should be noted that if $\|f^*\|_\infty < B$, the generalization error of the clipped estimator $\check{f}_\Lambda$ is not greater than that of the original estimator $\hat{f}_\Lambda$,

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

because $|\check{f}_\Lambda(x) - f^*(x)| \le |\hat{f}_\Lambda(x) - f^*(x)|$ 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,

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

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

$$\begin{aligned}\Theta_n = &\big\{(\lambda_1, \lambda_2, \lambda_3) \mid \lambda_1, \lambda_3 \in \Gamma_n, \lambda_2 = \lambda_1 \lambda_3^{1/2}\big\} \\ &\cup \big\{(\lambda_1, \lambda_2, \lambda_3) \mid \lambda_1, \lambda \in \Gamma_n, \lambda_2 = \lambda_1 \lambda^{1/2}, \lambda_3 = 0\big\}.\end{aligned}$$

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

THEOREM 5. *Suppose Assumptions 1–4 are satisfied. Assume* R_{1,f^*}, $R_{2,g^*} \ge 1$, $\beta_{b_2} \ge \beta_{b_1} \ge C_2$ *and* $\|f_m^*\|_{\mathcal{H}_m}, \|g_m^*\|_{\mathcal{H}_m} \le C_3$ *with some constants* $C_2, C_3 > 0$, *and suppose* n *satisfies* $\frac{\log(M)}{\sqrt{n}} \le 1$ *and*

$$\frac{\tilde{C}_1}{\beta_{b_1}^2} \psi_s \sqrt{n} \xi_n(\lambda_{(1)})^2 d \le 1 \quad \text{and} \quad \frac{\tilde{C}_1}{\beta_{b_2}^2} \psi_s \sqrt{n} \xi_n(\lambda_{(2)})^2 d \le 1,$$

where $\lambda_{(1)} = d^{1/(1+q+s)} n^{-1/(1+q+s)} R_{2,g^*}^{-2/(1+q+s)}$, $\lambda_{(2)} = d^{(1-s)/(1+s)} n^{-1/(1+s)} \times R_{1,f^*}^{-2/(1+s)}$ *and* $\tilde{C}_1$ *is the constant introduced in the statement of Theorem 2. Then there exist a universal constant* $\tilde{C}_4$ *and a constant* $\tilde{C}_3$ *depending on* s, c, L, C_1, C_2, C_3 *such that*

$$\begin{aligned}&\|\check{f}_{\Lambda_{D_{\mathrm{te}}}} - f^*\|_{L_2(\Pi)}^2 \\ &\quad \le \tilde{C}_3 \Bigg(d^{(1-s)/(1+s)} n^{-1/(1+s)} R_{1,f^*}^{2s/(1+s)} \\ &\qquad \wedge d^{(1+q)/(1+q+s)} n^{-(1+q)/(1+q+s)} R_{2,g^*}^{2s/(1+q+s)} + \frac{d \log(M)}{n} \Bigg) \eta(t)^2 \\ &\qquad + \tilde{C}_4 \frac{B^2(\tau + \log(1+n))}{n},\end{aligned}$$

with probabitlity $1 - 2\exp(-t) - \exp(-\tau) - \frac{2}{M}$, *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 5, the estimator $\check{f}_{\Lambda_{D_{\mathrm{te}}}}$ 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, the optimal rate is almost attainable [at the cost of the term $\frac{B^2 \log(1+n)}{n}$] by a simple executable algorithm.

5. Comparison with existing bounds. In this section, we compare our bound with the existing bounds. Roughly speaking, the difference between 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 rate 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 $\frac{1}{d}\sum_{m\in I_0}(\|g_m^*\|_{\mathcal{H}_m}^2/\|f_m^*\|_{\mathcal{H}_m}^2) \le C$. Under these rather strong conditions, they showed the bound

$$d^{(1-s)/(1+s)} n^{-1/(1+s)} + \frac{d\log(M)}{n}.$$

Our convergence rate (8) 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 $n^{-2/(2+s)}$ instead of $n^{-1/(1+s)}$ with respect to n.

The second bound was given by Meier, van de Geer and Bühlmann (2009), which shows

$$\left(\frac{\log(M)}{n}\right)^{1/(1+s)} (d + R_{2,f^*}^2)$$

for elastic-net regularization under the condition $q = 0$. Their bound almost achieves the minimax rate on the ℓ_∞-mixed-norm ball except the $\log(M)$ factor. Compared with our bound (9), their bound has the additional $\log(M)$ factor and the term with respect to d and R_{2,f^*} is larger than $d^{1/(1+s)} R_{2,f^*}^{2s/(1+s)}$ in our learning rate of elastic-net MKL because Young's inequality yields

$$d^{1/(1+s)} R_{2,f^*}^{2s/(1+s)} \le \frac{1}{1+s} d + \frac{s}{1+s} R_{2,f^*}^2 \le d + R_{2,f^*}^2.$$

Moreover, our result for elastic-net MKL covers all $0 \le q \le 1$.

Most recently, Koltchinskii and Yuan (2010) presented the bound

$$n^{-1/(1+s)}(d + R_{1,f^*}) + \frac{d\log(M)}{n}$$

for L_1-MKL and $q = 0$. Their bound achieves the minimax rate on the ℓ_∞-mixed-norm ball, but is looser than our bound (8) of L_1-MKL because, by Young's inequality, we have

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

In fact, their bound is $d^{2s/(1+s)}$ times slower than ours if the ground truth is *inhomogeneous*. To see this, suppose $\|f_m^*\|_{\mathcal{H}_m} = m^{-1}$ ($m \in I_0 = \{1, \ldots, d\}$) and $f_m^* = 0$ (otherwise). Then their bound is $n^{-1/(1+s)} d + \frac{d \log(M)}{n}$, while our bound for L_1-MKL is $n^{-1/(1+s)} d^{(1-s)/(1+s)} + \frac{d \log(M)}{n}$. Moreover, their formulation of L_1-MKL is slightly different from ours. In their formulation, there are additional constraints such that $\|f_m\|_{\mathcal{H}_m} \leq R_m$ ($\forall m$) with some constants R_m in the optimization problem described in equation (1). Due to these constraints, their formulation is a bit different from the practically used one (in practice, we do not usually impose such constrains). Instead, our analysis requires an additional assumption on the sup-norm (Assumption 4) to control the discrepancy between the empirical and population means of the square of an element in RKHS, $\frac{1}{n} \sum_{i=1}^n f_m^2(x_i) - \mathrm{E}[f_m^2]$ ($f_m \in \mathcal{H}_m$). In addition, 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, Wainwright and Yu (2012)]. Our analysis omits the global boundedness by utilizing the sup-norm assumption (Assumption 4).

All of 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 covered.

6. Discussion about adaptivity of ℓ_1-regularization. In this section, we discuss the issue, "is it really true that ℓ_1-regularization cannot possess adaptivity to the smoothness?" According to Theorem 2 and the following discussion, the convergence rate of L_1-MKL does not have dependency on the smoothness of the true function. However, this is just an upper bound. Thus, there is still possibility that L_1-MKL can make use of the smoothness of the true function. We give some remarks about this issue.

According to our analysis, it is difficult to improve the bound of Theorem 2 without any additional assumptions. On the other hand, it is possible to show this if we may assume some additional conditions.

A technical reason that makes it difficult to show adaptivity of L_1-MKL is that the ℓ_1-regularization is not differentiable at 0. Indeed, the sub-gradient of $\|f_m\|_{\mathcal{H}_m}$ is $f_m / \|f_m\|_{\mathcal{H}_m}$ if $f_m \neq 0$, and compared with that of $\|f_m\|_{\mathcal{H}_m}^2$ (which is f_m), there is a difference of a factor $1/\|f_m\|_{\mathcal{H}_m}$. This makes it difficult to control the behavior of the estimator around 0. To avoid this difficulty, we assume that the estimator $\hat{f}_m$ is bounded below as follows.

ASSUMPTION 5 (Lower bound assumption). There exist constants $h_m > 0$ ($m \in I_0$) such that

$$\|\hat{f}_m\|_{\mathcal{H}_m} \geq h_m \qquad (\forall m \in I_0), \tag{A5}$$

with probability $1 - p_n$.

We will give a justification of this assumption later (Lemma 7). If we admit this assumption, we have the following convergence bound. Define

$$\hat{R}_{2,g^*} := \Bigg(\sum_{m \in I_0} \frac{\|g_m^*\|_{\mathcal{H}_m}^2}{h_m} \Bigg)^{1/2},$$

$$b_3 := 32 \Bigg(1 + \frac{\sqrt{d} \max_{m \in I_0} (\|g_m^*\|_{\mathcal{H}_m} / h_m)}{\hat{R}_{2,g^*}} \Bigg).$$

THEOREM 6. *Suppose Assumptions 1–5 are satisfied, and* $\|g_m^*\|_{\mathcal{H}_m} \leq C$ *for all* $m \in I_0$. *Set*

$$\lambda = d^{1/(1+q+s)} n^{-1/(1+q+s)} \hat{R}_{2,g^*}^{-2/(1+q+s)}.$$

Moreover we set $\lambda_1^{(n)}$, $\lambda_2^{(n)}$ *and* $\lambda_3^{(n)}$ *as* $\lambda_1^{(n)} = 2\psi_s \eta(t) \xi_n(\lambda)$, $\lambda_2^{(n)} = \max\{\lambda \eta(t), \lambda_1^{(n)} \lambda^{1/2}\}$, $\lambda_3^{(n)} = 0$ *where* ψ_s *is same as Theorem 2. Similarly define* $\lambda_1^{(n)}(t')$, $\lambda_2^{(n)}(t')$ *corresponding to some fixed* t', *and* $\tilde{\lambda} = (\lambda_2^{(n)}(t') / \lambda_1^{(n)}(t'))^2$. *Then there exist constants* $\tilde{C}_3$, $\tilde{C}_3'$, $\tilde{C}_4$ *depending on* s, c, L, C_1, C, b_3, t' *such that for all* n *satisfying* $\frac{\log(M)}{\sqrt{n}} \leq 1$ *and*

$$\frac{\tilde{C}_3}{\beta_{b_3}^2} \psi_s \sqrt{n} \xi_n^2(\lambda) d \leq 1, \qquad \tilde{C}_3' \psi_s \sqrt{n} \xi_n^2(\tilde{\lambda}) \tilde{\lambda} d \leq \lambda_2^{(n)}(t'), \tag{14}$$

we have that

$$\begin{aligned} &\|\hat{f} - f^*\|_{L_2(\Pi)}^2 \\ &\quad \leq \frac{\tilde{C}_4}{\beta_{b_3}^2} \Bigg(d^{(1+q)/(1+q+s)} n^{-(1+q)/(1+q+s)} \hat{R}_{2,g^*}^{2s/(1+q+s)} + \frac{d \log(M)}{n} \Bigg) \eta(t)^2, \end{aligned} \tag{15}$$

with probability $1 - \exp(-t) - \exp(-t') - 2/M - p_n$.

The proof of Theorem 6 can be found in Section S.4 of the supplementary material [Suzuki and Sugiyama (2013)]. The theorem shows that with the rather strong assumption (Assumption 5), we can show that L_1-MKL also possesses adaptivity to the smoothness. Bound (15) is close to the minimax optimal rate on the ℓ_2-mixed-norm ball where $\hat{R}_{2,g^*}$ appears instead of R_{2,g^*}. Here we observe that h_m

appears in the denominator in $\hat{R}_{2,g^*}$. Therefore, for small h_m, $\hat{R}_{2,g^*}$ is larger than R_{2,g^*}, which can make bound (15) larger than that of elastic-net MKL. This is due to the indifferentiability of ℓ_1-regularization as explained above.

Next, we give a justification of Assumption 5.

LEMMA 7. *If* $\|\hat{f}_m - f_m^*\|_{L_2(\Pi)} \to 0$ *in probability, then*

$$P\left(\|\hat{f}_m\|_{\mathcal{H}_m} \geq \frac{\|f_m^*\|_{\mathcal{H}_m}}{2}\right) \to 1.$$

PROOF. On the basis of decomposition (2) of the kernel function, we write $f_m^* = \sum_{j=1}^{\infty} a_{j,m}\phi_{j,m}$ and $\hat{f}_m = \sum_{j=1}^{\infty} \hat{a}_{j,m}\phi_{j,m}$. Then we have that $\|f_m^*\|_{\mathcal{H}_m}^2 = \sum_{j=1}^{\infty} \mu_{j,m}^{-1} a_{j,m}^2$. Now we define $J_{f_m^*}$ to be a finite number such that $\sqrt{\sum_{j=1}^{J_{f_m^*}} \mu_{j,m}^{-1} a_{j,m}^2} \geq \frac{3}{4}\|f_m^*\|_{\mathcal{H}_m}$. Noticing that $o_p(1) \geq \|\hat{f}_m - f_m^*\|_{L_2(\Pi)}^2 = \sum_{j=1}^{\infty}(a_{j,m} - \hat{a}_{j,m})^2 \geq \sum_{j=1}^{J_{f_m^*}}(a_{j,m} - \hat{a}_{j,m})^2$, we have that

$$\begin{aligned}
\|\hat{f}_m\|_{\mathcal{H}_m} &= \sqrt{\sum_{j=1}^{J_{f_m^*}} \mu_{j,m}^{-1}\hat{a}_{j,m}^2 + \sum_{j=J_{f_m^*}+1}^{\infty} \mu_{j,m}^{-1}\hat{a}_{j,m}^2} \\
&\geq \sqrt{\sum_{j=1}^{J_{f_m^*}} \mu_{j,m}^{-1}\hat{a}_{j,m}^2} \\
&\geq \sqrt{\sum_{j=1}^{J_{f_m^*}} \mu_{j,m}^{-1} a_{j,m}^2} - \sqrt{\sum_{j=1}^{J_{f_m^*}} \mu_{j,m}^{-1}(a_{j,m} - \hat{a}_{j,m})^2} \\
&\geq \frac{3}{4}\|f_m^*\|_{\mathcal{H}_m} - \mu_{J_{f_m^*}}^{-1/2} \sqrt{\sum_{j=1}^{J_{f_m^*}}(a_{j,m} - \hat{a}_{j,m})^2} = \frac{3}{4}\|f_m^*\|_{\mathcal{H}_m} - o_p(1).
\end{aligned}$$

This gives the assertion. □

One can see from the proof that the convergence rate in Lemma 7 depends on f_m^*. If d is sufficiently small, we observe that the proof of Theorem 2 gives that $\|f_m^* - \hat{f}_m\|_{L_2(\Pi)} \xrightarrow{p} 0$ $(m \in I_0)$. In this situation, if we set $h_m = \|f_m^*\|_{\mathcal{H}_m}/2$, $\|f_m^*\|_{\mathcal{H}_m} \geq h_m$ $(m \in I_0)$ is satisfied with high probability for sufficiently large n.

The above discussion seems a proper justification to support the adaptivity of ℓ_1-regularization. However, we would like to remark the following two concerns about the discussion. First, in a situation where d increases as the number of samples increases, it is hardly expected that $\|f_m^*\|_{\mathcal{H}_m} > c$ with some positive constant c. It is more natural to suppose that $\min_{m \in I_0} \|f_m^*\|_{\mathcal{H}_m} \to 0$ as d increases. In

that situation, $\hat{R}_{2,g^*}$ becomes much larger as d increases. Second, since T_m is not invertible, $\|g_m^*\|_{\mathcal{H}_m}/\|f_m^*\|_{\mathcal{H}_m}$ is not bounded. Thus for $h_m = \|f_m^*\|_{\mathcal{H}_m}/2$, we have no guarantee that $\hat{R}_{2,g^*}$ is reasonably small so that the convergence bound (15) is meaningful. Both of these two concerns are caused by the indiffirentiability of ℓ_1-regularization at 0. Moreover these concerns are specific to high-dimensional situations. If $d = M = 1$ (or d and M are sufficiently small), then we do not need to worry about such issues.

We have shown that in a restrictive situation, ℓ_1-regularization can possess adaptivity to the smoothness of the true function and achieve a near minimax optimal rate on the ℓ_2-mixed-norm ball. It is a future work to clarify whether the lower bounded assumption (Assumption 5) is a necessary condition or not.

7. Conclusion. We have presented a new learning rate of both L_1-MKL and elastic-net MKL, which is tighter 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 have also shown that a procedure like cross validation gives the optimal choice of the parameters. We have discussed a relation between the regularization and the convergence rate. Our theoretical analysis suggests that there is a trade-off between the sparsity and the smoothness; that is, if the true function is sufficiently smooth, elastic-net regularization is preferred; otherwise, ℓ_1-regularization is preferred. This theoretical insight supports the recent experimental results [Cortes, Mohri and Rostamizadeh (2009b), Kloft et al. (2009), Tomioka and Suzuki (2009)] such that intermediate regularization between ℓ_1 and ℓ_2 often shows favorable performances.

APPENDIX A: EVALUATION OF ENTROPY 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

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

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

$$\mathcal{H}_{m,\beta} := \left\{ \sum_{k=1}^{\infty} b_k \phi_{k,m} \;\middle|\; \sum_{k=1}^{\infty} \mu_{k,m}^{-\beta} b_k^2 < \infty \right\},$$

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 $\mathcal{H}_{m,1} = \mathcal{H}_m$; see Theorem 4.51 of Steinwart

and Christmann (2008). Here we define, for $R > 0$,

$$\mathcal{H}_m^q(R) := \{ f_m = T_m^{q/2} g_m \mid g_m \in \mathcal{H}_m, \|g_m\|_{\mathcal{H}_m} \leq R \}. \tag{16}$$

Then we obtain the following lemma.

LEMMA 8. *$\mathcal{H}_m^q(1)$ is equivalent to the unit ball of $\mathcal{H}_{m,1+q}$: $\mathcal{H}_m^q(1) = \{f_m \in \mathcal{H}_{m,1+q} \mid \|f_m\|_{\mathcal{H}_{m,1+q}} \leq 1\}$.*

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 $f_m = T_m^{q/2} g_m$ and $\|g_m\|_{\mathcal{H}_m} \leq 1$. Thus $g_m = (T_m^{q/2})^{-1} f_m = \sum_{k=1}^\infty \mu_{k,m}^{-q/2} \langle f_m, \phi_{k,m} \rangle_{L_2(\Pi)} \phi_{k,m}$ and $1 \geq \|g_m\|_{\mathcal{H}_m} = \sum_{k=1}^\infty \mu_{k,m}^{-1} \langle g_m, \phi_{k,m} \rangle_{L_2(\Pi)}^2 = \sum_{k=1}^\infty \mu_{k,m}^{-(1+q)} \langle f_m, \phi_{k,m} \rangle_{L_2(\Pi)}^2$. Therefore, f_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, Hush and Scovel (2009) gives an upper bound of the entropy number of $\mathcal{H}_{m,\beta}$ as

$$e_i\big(\mathcal{H}_{m,\beta} \to L_2(\Pi)\big) \leq C i^{-\beta/(2s)},$$

where C is a constant depending on c, s, β. This inequality with $\beta = 1$ corresponds to equation 3. Moreover, substituting $\beta = 1 + q$ into the above equation, we have

$$e_i\big(\mathcal{H}_{m,\beta} \to L_2(\Pi)\big) \leq C i^{-(1+q)/(2s)}. \tag{17}$$

APPENDIX B: PROOF OF LEMMA 1

PROOF OF LEMMA 1. For $J = I^c$, we have

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

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

Acknowledgments. The authors would like to thank Ryota Tomioka, Alexandre B. Tsybakov, Martin Wainwright and Garvesh Raskutti for suggestive discussions.

SUPPLEMENTARY MATERIAL

Supplementary material for: Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness (DOI: 10.1214/13-AOS1095SUPP; .pdf). Due to space constraints, we have moved the proof of the main theorem to a supplementary document [Suzuki and Sugiyama (2013)].

REFERENCES

ARGYRIOU, A., HAUSER, R., MICCHELLI, C. A. and PONTIL, M. (2006). A DC-programming algorithm for kernel selection. In *The 23st International Conference on Machine Learning* (W. W. Cohen and A. Moore, eds.). ACM, New York.

BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

BACH, F. R. (2009). Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems* **21** (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 105–112. Curran Associates, Red Hook, NY.

BACH, F. R., LANCKRIET, G. and JORDAN, M. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *The 21st International Conference on Machine Learning* 41–48. ACM, New York.

BENNETT, C. and SHARPLEY, R. (1988). *Interpolation of Operators*. *Pure and Applied Mathematics* **129**. Academic Press, Boston, MA. MR0928802

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.

CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368. MR2335249

CHAPELLE, O., VAPNIK, V., BOUSQUET, O. and MUKHERJEE, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning* **46** 131–159.

CORTES, C., MOHRI, M. and ROSTAMIZADEH, A. (2009a). Learning non-linear combinations of kernels. In *Advances in Neural Information Processing Systems* **22** (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 396–404. Curran Associates, Red Hook, NY.

CORTES, C., MOHRI, M. and ROSTAMIZADEH, A. (2009b) L_2 regularization for learning kernels. In *The 25th Conference on Uncertainty in Artificial Intelligence* (*UAI* 2009) (J. Bilmes and A. Ng, eds.). AUAI Press, Corvallis.

FERREIRA, J. C. and MENEGATTO, V. A. (2009). Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations Operator Theory* **64** 61–81. MR2501172

KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95. MR0290013

KLOFT, M. and BLANCHARD, G. (2012). On the convergence rate of ℓ_p-norm multiple kernel learning. *J. Mach. Learn. Res.* **13** 2465–2501. MR2973607

KLOFT, M., RÜCKERT, U. and BARTLETT, P. L. (2010). A unifying view of multiple kernel learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases* (*ECML/PKDD*) (J. L. Balcázar, F. Bonchi, A. Gionis and M. Sebag, eds.). *Lecture Notes in Computer Science* **6322** 66–81. Springer, Berlin.

KLOFT, M., BREFELD, U., SONNENBURG, S., LASKOV, P., MÜLLER, K. R. and ZIEN, A. (2009). Efficient and accurate ℓ_p-norm multiple kernel learning. In *Advances in Neural Information Processing Systems* **22** (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 997–1005. Curran Associates, Red Hook, NY.

KOLTCHINSKII, V. and YUAN, M. (2008). Sparse recovery in large ensembles of kernel machines. In *Proceedings of the Annual Conference on Learning Theory* (R. Servedio and T. Zhang, eds.) 229–238. Omnipress, Madison, WI.

KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695. MR2766864

LANCKRIET, G., CRISTIANINI, N., GHAOUI, L. E., BARTLETT, P. and JORDAN, M. (2004). Learning the kernel matrix with semi-definite programming. *J. Mach. Learn. Res.* **5** 27–72.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

MICCHELLI, C. A. and PONTIL, M. (2005). Learning the kernel function via regularization. *J. Mach. Learn. Res.* **6** 1099–1125. MR2249850

ONG, C. S., SMOLA, A. J. and WILLIAMSON, R. C. (2005). Learning the kernel with hyperkernels. *J. Mach. Learn. Res.* **6** 1043–1071. MR2249848

RASKUTTI, G., WAINWRIGHT, M. and YU, B. (2009). Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness. In *Advances in Neural Information Processing Systems* **22** (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1563–1570. Curran Associates, Red Hook, NY.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. MR2913704

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

SHAWE-TAYLOR, J. (2008). Kernel learning for novelty detection. In *NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*.

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, New York.

SREBRO, N. and BEN-DAVID, S. (2006). Learning bounds for support vector machines with learned kernels. In *Learning Theory. Lecture Notes in Computer Science* **4005** 169–183. Springer, Berlin. MR2280605

STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

STEINWART, I., HUSH, D. and SCOVEL, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.) 79–93. Omnipress, Madison, WI.

SUZUKI, T. (2011a). Unifying framework for fast learning rate of non-sparse multiple kernel learning. In *Advances in Neural Information Processing Systems* **24** (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, eds.) 1575–1583. Curran Associates, Red Hook, NY.

SUZUKI, T. (2011b). Fast learning rate of non-sparse multiple kernel learning and optimal regularization strategies. Available at arXiv:1111.3781.

SUZUKI, T. and SUGIYAMA, M. (2013). Supplement to "Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness." DOI:10.1214/13-AOS1095.

TOMIOKA, R. and SUZUKI, T. (2009). Sparsity-accuracy trade-off in MKL. In *NIPS 2009 Workshop: Understanding Multiple Kernel Learning Methods.*

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

VARMA, M. and BABU, B. R. (2009). More generality in efficient multiple kernel learning. In *The 26th International Conference on Machine Learning* (L. Bottou and M. Littman, eds.) 1065–1072. Omnipress, Madison, WI.

YING, Y. and CAMPBELL, C. (2009). Generalization bounds for learning the kernel. In *Proceedings of the Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.). Omnipress, Madison, WI.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITY OF TOKYO
7-3-1 HONGO, BUNKYO-KU
TOKYO
JAPAN
E-MAIL: s-taiji@stat.t.u-tokyo.ac.jp

DEPARTMENT OF COMPUTER SCIENCE
GRADUATE SCHOOL OF INFORMATION SCIENCE AND ENGINEERING
TOKYO INSTITUTE OF TECHNOLOGY
2-12-1 O-OKAYAMA, MEGURO-KU
TOKYO
JAPAN
E-MAIL: sugi@cs.titech.ac.jp