Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2013, Article ID 715275, 7 pages
http://dx.doi.org/10.1155/2013/715275

*Research Article*

# Least Square Regularized Regression for Multitask Learning

**Yong-Li Xu,[1] Di-Rong Chen,[2] and Han-Xiong Li[3]**

[1] *Department of Mathematics, Beijing University of Chemical Technology, Beijing 100029, China*
[2] *Department of Mathematics, Beijing University of Aeronautics and Astronautics, Beijing 100091, China*
[3] *Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong*

Correspondence should be addressed to Yong-Li Xu; xuyongli2312@sina.com

Received 11 October 2013; Accepted 13 November 2013

Academic Editor: Yiming Ying



The study of multitask learning algorithms is one of very important issues. This paper proposes a least-square regularized regression algorithm for multi-task learning with hypothesis space being the union of a sequence of Hilbert spaces. The algorithm consists of two steps of selecting the optimal Hilbert space and searching for the optimal function. We assume that the distributions of different tasks are related to a set of transformations under which any Hilbert space in the hypothesis space is norm invariant. We prove that under the above assumption the optimal prediction function of every task is in the same Hilbert space. Based on this result, a pivotal error decomposition is founded, which can use samples of related tasks to bound excess error of the target task. We obtain an upper bound for the sample error of related tasks, and based on this bound, potential faster learning rates are obtained compared to single-task learning algorithms.

## 1. Introduction

Multitask learning [1] is a learning paradigm which seeks to improve the generalization performance of a learning task with the help of some other related tasks. This learning paradigm is inspired by human learning activities in that people often apply the knowledge gained from previous learning tasks to help learn a new task [2]. Different multitask learning algorithms have been designed, such as multitask support vector machine (SVM) [3], multitask feature learning [4, 5], multitask clustering approach [6], multitask structure learning [7], and multitask gradients learning [8].

Multitask learning can be formulated under two different settings: symmetric and asymmetric [9]. The symmetric multitask learning tries to improve the performance of all tasks simultaneously, and the objective of asymmetric multitask learning tries to improve the performance of some target task using information from related tasks. The asymmetric multitask learning is related to transfer learning [10]. The major difference is that the source tasks are still learned simultaneously in asymmetric multitask learning while they are learned independently in transfer learning [2]. Much experimental work has achieved the target that is improving the prediction performance of a learning task with the help of other related tasks [1, 11–15]. However, there has been relatively little progress on theoretical analysis for these results.

Baxter presented a general frame for model selection in multitask learning environment [16]. They showed that a hypothesis space that performs well on a sufficiently large number of training tasks will also perform well when learning novel tasks in the same environment. They proved that learning multiple tasks within an environment of related tasks can potentially give much better generalization than learning a single task. Ando and Zhang considered learning predictive structures on hypothesis spaces from multitask learning [17]. They presented a general framework in which the structural learning problem can be formulated and analyzed theoretically and related it to learning with unlabeled data. Ben-David and Borbely defined relatedness of tasks on the basis of similarity between the example generating distributions for classification [18], and then they gave precise conditions under which bounds guarantee generalization on the basis of smaller sample sizes than the standard single-task approach. Solnon et al. studied multitask regression, using penalization techniques [19]. They showed that the key element appearing for an optimal calibration is the covariance matrix of the

noise between the different tasks. They presented a new algorithm to estimate this covariance matrix and proved that this estimator converges towards the covariance matrix.

In this paper, we propose a least-square regularized regression algorithm for multitask learning with hypothesis space being the union of a sequence of Hilbert spaces. The relatedness of tasks is described by distributions that underlie these tasks and some property of the hypothesis space. We assume that the distributions are related by a set of transformations under which the norm of any Hilbert space in the hypothesis space is invariant. We design a multitask learning algorithm with two steps: firstly, samples of other related tasks are used to select an approximate optimal Hilbert space in the hypothesis space; secondly, in the optimal Hilbert space, we use standard least square regularized regression algorithm for the target task. It is proved that, under the above assumption, the optimal prediction function of every task is in the same Hilbert space. For error analysis, we decompose the excess error of prediction function in target task into regularization error and sample error in which the difference between error and empirical error of the prediction function in the target task is estimated by the average value of those in related tasks. This leads to a potential faster learning rate than that of standard regularized regression algorithm in single task.

The rest of the paper is organized as follows. In Section 2, we introduce some notions and definitions and then propose the least square regularized regression for multitask learning. In Section 3, we decompose excess error of target task into regularization error and sample error in which samples of other related tasks can be used to estimated difference between error and empirical error of the prediction function in target task. The main result is presented in Section 4. An upper bound for sample error of multiple tasks is given by Hoeffding's inequality and an estimation for covering number in multitask learning, and then, based on the upper bound, potential faster learning rates compared to single-task learning algorithms are obtained.

## 2. Preliminaries

To propose regularized regression algorithm for multitask learning, we introduce some definitions and notations. Let $(X, d)$ be a compact metric space and let $Y = [-M, M]$. Let $\rho$ be a probability distribution on $Z := X \times Y$. The regression function is defined as

$$f_\rho(x) = \int_Y y \, d\rho(y \mid x), \quad x \in X, \tag{1}$$

where $\rho(y \mid x)$ is the conditional probability measure at $x$ induced by $\rho$. Knowing a set of samples from the probability distribution $\rho$, our goal is to find a good approximation of $f_\rho$.

For multitask learning, we define the relatedness of probability distribution of multiple tasks.

*Definition 1.* For a function $f : X \to X$, let $f[P]$ be the probability distribution over $X \times Y$ defined by $f[\rho](T) = \rho(\{(f(x), b) \mid (x, b) \in T\})$, for $T \subseteq X \times Y$. Let $\mathcal{F}$ be a set of transformations $f : X \to X$ and let $\rho_1$ and $\rho_2$ be probability distributions over $X \times Y$. We say that $\rho_1$ and $\rho_2$ are $\mathcal{F}$-related if there exists some $f \in \mathcal{F}$ such that $\rho_1 = f[\rho_2]$ or $\rho_2 = f[\rho_1]$.

Then, we describe the relatedness of the transformation set $\mathcal{F}$ above and a hypothesis space.

*Definition 2.* Let $\mathcal{F}$ be a set of transformations $f : X \to X$ and let $\mathcal{H}$ be a set of functions $X \to Y$. We say that $\mathcal{F}$ acts as a group over $\mathcal{H}$, if,

(1) for every $f \in \mathcal{F}$ and every $h \in \mathcal{H}$, there holds $h \circ f \in \mathcal{H}$;

(2) for every $f, g \in \mathcal{F}$, the inverse transformation $f^{-1}$ and the composition $f \circ g$ are also members of $\mathcal{F}$.

*Definition 3.* Let $\mathcal{H}_\sigma$ be a Hilbert space with norm $\|\cdot\|_\sigma$ and let $\mathcal{F}$ act as a group over $\mathcal{H}_\sigma$. We say $\mathcal{H}_\sigma$ is norm invariant under $\mathcal{F}$, if, for any $h \in \mathcal{H}_\sigma$ and any $f \in \mathcal{F}$, there holds

$$\|h\|_\sigma = \|h \circ f\|_\sigma. \tag{2}$$

To explain the above definitions, we give an example.

*Example 4.* Let $X = \mathbb{R}^2$, $Y = [-M, M]$, $\Gamma = R$, and $K_\sigma(x, x') = \exp\{-\|x - x'\|_2^2/\sigma^2\}$, for $x, x' \in X$, and $\mathcal{H}_\sigma$ be the closure of linear span of the set $\{K_{\sigma,x} := K_\sigma(x, \cdot) : x \in X\}$ with inner product $\langle K_{\sigma,x}, K_{\sigma,x'}\rangle_\sigma = K_\sigma(x, x')$, for any $\sigma \in \Gamma$. Let the norm $\|\cdot\|_\sigma$ of functions in $\mathcal{H}_\sigma$ be induced by the inner product.

Assume $\mathcal{F}$ is a set of translation and rotation transformations on $\{\mathcal{H}_\sigma, \sigma \in \Gamma\}$: for any $f \in \mathcal{F}$ and $h \in \{\mathcal{H}_\sigma, \sigma \in \Gamma\}$, there holds $(h \circ f)(x) := h(x + x_f)$, for some $x_f \in X$, or $(h \circ f)(x) := h(A_f x)$, for

$$A_f = \begin{bmatrix} \cos\theta_f & -\sin\theta_f \\ \sin\theta_f & \cos\theta_f \end{bmatrix}, \tag{3}$$

where $\theta_f$ is an angle dependent on $f$. We can verify that $\mathcal{F}$ acts as a group over $\mathcal{H}_\sigma$ and $\mathcal{H}_\sigma$ is norm invariant under $\mathcal{F}$, for any $\sigma \in \Gamma$.

Now we introduce standard least-square regularized regression algorithm. We denote error and empirical error of a function $f : X \to Y$ with squared loss as follows. For a distribution $\rho$ on $X \times Y$, the error of $f$ is defined as

$$E^\rho(f) = \int_Z (f(x) - y)^2 d\rho. \tag{4}$$

It is well known that the regression function minimizes the error. Indeed,

$$\|f - f_\rho\|_{\rho_X}^2 = \mathcal{E}(f) - \mathcal{E}(f_\rho), \tag{5}$$

where $\rho_X$ is the marginal distribution of $\rho$ on $X$ and $\|f\|_{\rho_X}^2 = \int_X |f(x)|^2 d\rho_X$. The above difference is called the excess error of $f$, and for a sample set $S = \{(x_i, y_i), i = 1, 2, ..., m\}$, independently drawn according to $\rho$, the empirical error of $f$ is defined as

$$\widehat{E}^S(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2. \tag{6}$$

In this paper, we consider a sequence of related learning tasks. The goal is to use information of related tasks to improve learning performance of one special learning task. Let $\mathscr{F}$ be a transformation set and $\mathscr{H}_\sigma$ be a Hilbert space with norm $\|\cdot\|_\sigma$, for any $\sigma \in \Gamma$, where $\Gamma$ is an index set. We assume that there is $\kappa > 0$, such that

$$\|f\|_\infty \le \kappa \|f\|_\sigma. \tag{7}$$

Let $\rho_i$ denote the probability distribution of the $i$th task, for $i = 1, \ldots, n$. We assume $\{\rho_i\}_{i=1}^n$ is pair-wise $\mathscr{F}$-related, $\mathscr{F}$ acts as a group over $\mathscr{H}_\sigma$, and $\mathscr{H}_\sigma$ is norm invariant under $\mathscr{F}$, for any $\sigma \in \Gamma$. Let $S_i = \{(x_i^j, y_i^j)\}_{j=1}^m$ be samples independently drawn according to $\rho_i$. Since the $n$ tasks are related, we try to use samples $\mathbf{z} = \cup_{i=1}^n S_i$ of all the $n$ tasks to improve the learning performance of the target task.

Standard least-square regularized regression algorithm associated with $\{\mathscr{H}_\sigma : \sigma \in \Gamma\}$ for the $i$th single task is defined as the minimizer

$$f_{S_i} = \arg \min_{f \in \mathscr{H}_\sigma, \sigma \in \Gamma} \left\{ \widehat{E}^{S_i}(f) + \lambda \|f\|_\sigma^2 \right\}. \tag{8}$$

In the above optimization, there are two steps; firstly, for any fixed $\sigma \in \Gamma$, find the optimal function $f_{S_i,\sigma}$; secondly, find the global optimal function $f_S$, for $\sigma \in \Gamma$. Since the $n$ tasks are related, we try to use samples $\mathbf{z} = \cup_{i=1}^n S_i$ of all the $n$ tasks to improve the learning performance of the target task. Without loss of generality, we choose the first task as the target task.

Now, we propose least square regularized regression algorithm for multitask learning.

*Step 1.* Use samples of other $n-1$ tasks to select the approximate optimal $\mathscr{H}_{\widehat{\sigma}}$ as follows:

$$\widehat{\sigma} = \arg \min_{\substack{\sigma \in \Gamma, \\ h_2, \ldots, h_n \in \mathscr{H}_\sigma}} \sum_{i=2}^n \left\{ \widehat{E}^{S_i}(h_i) + \lambda \|h_i\|_\sigma^2 \right\}. \tag{9}$$

*Step 2.* In $\mathscr{H}_{\widehat{\sigma}}$, search for the approximation $\widehat{h}_1$ to $f_{\rho_1}$ as follows:

$$\left(\widehat{h}_1, \ldots, \widehat{h}_n\right) = \arg \min_{h_1, \ldots, h_n \in \mathscr{H}_{\widehat{\sigma}}} \sum_{i=1}^n \left\{ \widehat{E}^{S_i}(h_i) + \lambda \|h_i\|_{\widehat{\sigma}}^2 \right\}. \tag{10}$$

## 3. Error Decomposition

To estimate the bound of excess error $E^{\rho_1}(\widehat{h}_1) - E^{\rho_1}(f_{\rho_1})$, we introduce some notations. Let $\sigma_i^*$ be the optimal $\sigma \in \Gamma$ for the $i$th task,

$$\left(\sigma_i^*, \overline{h}_i^*\right) = \arg \min_{\substack{\sigma \in \Gamma, \\ h \in \mathscr{H}_\sigma}} \left\{ E^{P_i}(h) + \lambda \|h\|_\sigma^2 \right\}. \tag{11}$$

**Lemma 5.** *Let $\rho_i$ denote the probability distribution of the $i$th task for $i = 1, \ldots, n$, let $\mathscr{H}_\sigma$ be a Hilbert space with norm $\|\cdot\|_\sigma$ for $\sigma \in \Gamma$, and let $\mathscr{F}$ be a transformation set. Assume $\{\rho_i\}_{i=1}^n$ is pair-wise $\mathscr{F}$-related, $\mathscr{F}$ acts as a group over $\mathscr{H}_\sigma$, and $\mathscr{H}_\sigma$ is norm invariant under $\mathscr{F}$, for any $\sigma \in \Gamma$. Then $\sigma_i^*$ defined in* (11) *satisfies*

$$\sigma_1^* = \sigma_2^* = \cdots = \sigma_n^*. \tag{12}$$

*Proof.* By the assumption that $\rho_i$ and $\rho_j$ are $\mathscr{F}$-related, it is easy to show that

$$\inf_{h \in \mathscr{H}_\sigma} E^{\rho_i}(h) = \inf_{h \in \mathscr{H}_\sigma} E^{\rho_j}(h), \tag{13}$$

for any $i, j \in \{1, \ldots, n\}$ and $\sigma \in \Gamma$. Notice that $\mathscr{H}_\sigma$ is norm-invariant under $\mathscr{F}$ for any $\sigma \in \Gamma$. Then, there holds

$$\inf_{h \in \mathscr{H}_\sigma} \left\{ E^{\rho_i}(h) + \lambda \|h\|_\sigma^2 \right\} = \inf_{h \in \mathscr{H}_\sigma} \left\{ E^{\rho_j}(h) + \lambda \|h\|_\sigma^2 \right\}, \tag{14}$$

for any $\sigma \in \Gamma$. Then the lemma follows. □

To decompose the excess error, we give the following notations. Denote

$$\begin{aligned} \left(\overline{h}_1, \ldots, \overline{h}_n\right) &= \arg \min_{h_1, \ldots, h_n \in \mathscr{H}_{\widehat{\sigma}}} \sum_{i=1}^n \left\{ E^{\rho_i}(h_i) + \lambda \|h_i\|_{\widehat{\sigma}}^2 \right\}, \\ \left(\widehat{h}_1^*, \ldots, \widehat{h}_n^*\right) &= \arg \min_{h_1, \ldots, h_n \in \mathscr{H}_{\sigma^*}} \sum_{i=1}^n \left\{ \widehat{E}^{S_i}(h_i) + \lambda \|h_i\|_{\sigma^*}^2 \right\}. \end{aligned} \tag{15}$$

**Proposition 6.** *Let $\widehat{h}_i$ be defined in* (10). *Then under the assumption of Lemma 5, $E^{\rho_1}(\widehat{h}_1) - E^{\rho_1}(f_{\rho_1}) \le E^{\rho_1}(\widehat{h}_1) - E^{\rho_1}(f_{\rho_1}) + \lambda \|\widehat{h}_1\|_{\widehat{\sigma}}^2$ can be bounded by*

$$\begin{aligned} &\left\{ E^{\rho_1}\left(\overline{h}_1^*\right) - E^{\rho_1}\left(f_{\rho_1}\right) + \lambda \left\|\overline{h}_1^*\right\|_{\sigma^*}^2 \right\} + \left\{ E^{\rho_1}\left(\widehat{h}_1\right) - \widehat{E}^{S_1}\left(\widehat{h}_1\right) \right\} \\ &+ \left\{ \widehat{E}^{S_1}\left(\overline{h}_1\right) - E^{\rho_1}\left(\overline{h}_1\right) \right\} + \frac{1}{n-1} \sum_{i=2}^n \left\{ E^{\rho_i}\left(\overline{h}_i^*\right) - \widehat{E}^{S_i}\left(\overline{h}_i^*\right) \right\} \\ &+ \frac{1}{n-1} \sum_{i=2}^n \left\{ E^{\rho_i}\left(\widehat{h}_i\right) - \widehat{E}^{S_i}\left(\widehat{h}_i\right) \right\}. \end{aligned} \tag{16}$$

*Proof.* Write the regularization error as

$$\begin{aligned} &E^{\rho_1}\left(\widehat{h}_1\right) - E^{\rho_1}\left(f_{\rho_1}\right) + \lambda \left\|\widehat{h}_1\right\|_{\widehat{\sigma}}^2 \\ &= \left\{ E^{\rho_1}\left(\widehat{h}_1\right) - \widehat{E}^{S_1}\left(\widehat{h}_1\right) \right\} \\ &\quad + \left\{ \left( \widehat{E}^{S_1}\left(\widehat{h}_1\right) + \lambda \left\|\widehat{h}_1\right\|_{\widehat{\sigma}}^2 \right) - \left( \widehat{E}^{S_1}\left(\overline{h}_1\right) + \lambda \left\|\overline{h}_1\right\|_{\widehat{\sigma}}^2 \right) \right\} \\ &\quad + \left\{ \widehat{E}^{S_1}\left(\overline{h}_1\right) - E^{\rho_1}\left(\overline{h}_1\right) \right\} \\ &\quad + \left\{ \left( E^{\rho_1}\left(\overline{h}_1\right) + \lambda \left\|\overline{h}_1\right\|_{\widehat{\sigma}}^2 \right) - \left( E^{\rho_1}\left(\overline{h}_1^*\right) + \lambda \left\|\overline{h}_1^*\right\|_{\sigma^*}^2 \right) \right\} \\ &\quad + \left\{ E^{\rho_1}\left(\overline{h}_1^*\right) - E^{\rho_1}\left(f_{\rho_1}\right) + \lambda \left\|\overline{h}_1^*\right\|_{\sigma^*}^2 \right\}. \end{aligned} \tag{17}$$

By Lemma 5, for any $i, j \in \{1, \ldots, n\}$, there holds $E^{\rho_i}(\overline{h}_i) + \lambda\|\overline{h}_i\|^2_{\widehat{\sigma}} = E^{\rho_j}(\overline{h}_j) + \lambda\|\overline{h}_j\|^2_{\widehat{\sigma}}$. Then we obtain

$$
\begin{aligned}
&E^{\rho_1}\left(\overline{h}_1\right) + \lambda\left\|\overline{h}_1\right\|^2_{\widehat{\sigma}} \\
&= \frac{1}{n-1}\sum_{i=2}^{n}\left\{E^{\rho_i}\left(\overline{h}_i\right) + \lambda\left\|\overline{h}_i\right\|^2_{\widehat{\sigma}}\right\} \\
&= \frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(E^{\rho_i}\left(\overline{h}_i\right) + \lambda\left\|\overline{h}_i\right\|^2_{\widehat{\sigma}}\right) - \left(E^{\rho_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right)\right\} \\
&\quad + \frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(E^{\rho_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right) - \left(\widehat{E}^{S_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right)\right\} \\
&\quad + \frac{1}{n-1}\sum_{i=2}^{n}\left\{\widehat{E}^{S_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right\}, \\
&E^{\rho_1}\left(\overline{h}^*_1\right) + \lambda\left\|\overline{h}^*_1\right\|^2_{\sigma^*} \\
&= \frac{1}{n-1}\sum_{i=2}^{n}\left\{E^{\rho_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right\} \\
&= \frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(E^{\rho_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right)\right. \\
&\qquad \left. - \left(\widehat{E}^{S_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right)\right\} \\
&\quad + \frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(\widehat{E}^{S_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right)\right. \\
&\qquad \left. - \left(\widehat{E}^{S_i}\left(\widehat{h}^*_i\right) + \lambda\left\|\widehat{h}^*_i\right\|^2_{\sigma^*}\right)\right\} \\
&\quad + \frac{1}{n-1}\sum_{i=2}^{n}\left\{\widehat{E}^{S_i}\left(\widehat{h}^*_i\right) + \lambda\left\|\widehat{h}^*_i\right\|^2_{\sigma^*}\right\}.
\end{aligned}
\tag{18}
$$

By the definition, we have that

$$
\begin{aligned}
&\frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(E^{\rho_i}\left(\overline{h}_i\right) + \lambda\left\|\overline{h}_i\right\|^2_{\widehat{\sigma}}\right) - \left(E^{\rho_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right)\right\} \le 0, \\
&\frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(\widehat{E}^{S_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right) - \left(\widehat{E}^{S_i}\left(\widehat{h}^*_i\right) + \lambda\left\|\widehat{h}^*_i\right\|^2_{\sigma^*}\right)\right\} \\
&\quad \ge 0, \\
&\qquad \frac{1}{n-1}\sum_{i=2}^{n}\left\{\widehat{E}^{S_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right\} \\
&\qquad\quad \le \frac{1}{n-1}\sum_{i=2}^{n}\left\{\widehat{E}^{S_i}\left(\widehat{h}^*_i\right) + \lambda\left\|\widehat{h}^*_i\right\|^2_{\sigma^*}\right\}.
\end{aligned}
\tag{19}
$$

Therefore, we have

$$
\begin{aligned}
&\left(E^{\rho_1}\left(\overline{h}_1\right) + \lambda\left\|\overline{h}_1\right\|^2_{\widehat{\sigma}}\right) - \left(E^{\rho_1}\left(\overline{h}^*_1\right) + \lambda\left\|\overline{h}^*_1\right\|^2_{\sigma^*}\right) \\
&\le \frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(E^{\rho_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right) - \left(\widehat{E}^{S_i}\left(\widehat{h}_i\right) + \lambda\left\|\widehat{h}_i\right\|^2_{\widehat{\sigma}}\right)\right\} \\
&\quad + \frac{1}{n-1}\sum_{i=2}^{n}\left\{\left(E^{\rho_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right)\right. \\
&\qquad \left. - \left(\widehat{E}^{S_i}\left(\overline{h}^*_i\right) + \lambda\left\|\overline{h}^*_i\right\|^2_{\sigma^*}\right)\right\}.
\end{aligned}
\tag{20}
$$

Then, the proposition follows. □

In (16), there are 5 terms. The first term is called regularization error which depends on the approximation ability of hypothesis space to $f_{\rho_1}$. The estimation of this term has been discussed in [20] for reproducing kernel Hilbert space with Gaussian kernel with flexible variances.

The other four terms are called sample error. In the second and third terms, $\widehat{h}_1$ and $\overline{h}_1$ are selected from $\mathcal{H}_{\widehat{\sigma}}$ which is dependent on $S_i$, for $i = 2, \ldots, n$, and is independent of $S_1$. Therefore, when we take expectation of $S_1$, $\mathcal{H}_{\widehat{\sigma}}$ can be seen as a fixed function space. Consequently, these two terms can be estimated with the same method in the proof of Propositions 2.1 and 3.1 in [21]. In the fourth term, $\overline{h}^*_i$ is a fixed function, for $i = 1, \ldots, n$. Therefore, this term can also be estimated as in the proof of Proposition 2.1 in [21]. The last term is more difficult to deal with because $\widehat{h}_i$, for $i = 2, \ldots, n$, can not be considered in $\mathcal{H}_\sigma$, for any fixed $\sigma \in \Gamma$. Consequently, when sample number $m \to \infty$, the convergence rate of the sample error depends on that of the last term. Therefore, in the following section, we focus on the estimation for the bound of the last term in (16).

## 4. Error Analysis

In this section, we estimate the bound of the last term in (16). To bound this term, we have to estimate capacity of $\{\mathcal{H}_\sigma : \sigma \in \Gamma\}$. Here, the capacity is measured by the covering number.

*Definition 7.* For a subset $\mathcal{F}$ of a metric space $(X, d)$ and $\eta > 0$, the covering number $\mathcal{N}(\mathcal{F}, \eta, d)$ is defined to be the minimal integer $l \in \mathbb{N}$ such that there exist $l$ disks with radius $\eta$ covering $\mathcal{F}$.

For $n \in \mathbb{N}$ and $\mathcal{H}_{\sigma,R} = \{f \in \mathcal{H}_\sigma, \|f\|_\sigma \le R\}$, define

$$
\mathcal{H}^n_{\sigma,R} = \left\{f = (f_1, \ldots, f_n) : f_i \in \mathcal{H}_{\sigma,R}, i = 1, \ldots, n\right\}. \tag{21}
$$

For $f \in \mathcal{H}^n_{\sigma_1,R}$ and $g \in \mathcal{H}^n_{\sigma_2,R}$, define

$$
\begin{aligned}
l_{\infty,n}(f, g) &= \frac{1}{n}\sum_{i=1}^{n}\|f_i - g_i\|_\infty, \\
d_\infty\left(f, \mathcal{H}^n_{\sigma_2,R}\right) &= \inf_{g \in \mathcal{H}^n_{\sigma_2,R}} l_{\infty,n}(f, g).
\end{aligned}
\tag{22}
$$

Then, define the distance from $\mathcal{H}^n_{\sigma_1,R}$ to $\mathcal{H}^n_{\sigma_2,R}$ as

$$\widetilde{d}_\infty\left(\mathcal{H}^n_{\sigma_1,R},\mathcal{H}^n_{\sigma_2,R}\right)=\sup_{f\in\mathcal{H}^n_{\sigma_1,R}} d_\infty\left(f,\mathcal{H}^n_{\sigma_2,R}\right). \tag{23}$$

Let $\mathcal{N}_R(\varepsilon,\Gamma,\widetilde{d}_\infty)$ denote the minimal integer $l\in\mathbb{N}$ such that there exist $l$ parameters $\sigma_1,\ldots,\sigma_l\in\Gamma$, such that

$$\sup_{\sigma\in\Gamma}\min_{i\in\{1,\ldots,l\}}\widetilde{d}_\infty\left(\mathcal{H}^n_{\sigma,R},\mathcal{H}^n_{\sigma_i,R}\right)\le\varepsilon. \tag{24}$$

Then, for $\mathcal{H}^n_{\Gamma,R}=\{f=(f_1,\ldots,f_n):f_i\in\mathcal{H}_{\sigma,R},i=1,\ldots,n,\sigma\in\Gamma\}$, we have the following lemma.

**Lemma 8.** *Consider the following:*

$$\begin{aligned}&\mathcal{N}\left(2\varepsilon,\mathcal{H}^n_{\Gamma,R},l_{\infty,n}\right)\\&\quad\le\mathcal{N}_R\left(\varepsilon,\Gamma,\widetilde{d}_\infty\right)\cdot\left(\sup_{\sigma\in\Gamma}\mathcal{N}\left(\varepsilon,\mathcal{H}_{\sigma,R},l_{\infty,1}\right)\right)^n.\end{aligned} \tag{25}$$

*Proof.* For any $f\in\mathcal{H}^n_{\Gamma,R}$, there is $\sigma\in\Gamma$ such that $f\in\mathcal{H}^n_{\sigma,R}$. By the definition of $\mathcal{N}_R(\varepsilon,\Gamma,\widetilde{d}_\infty)$, there is $\overline{\sigma}\in\{\sigma_1,\sigma_2,\ldots,\sigma_{\mathcal{N}_R(\varepsilon,\Gamma,\widetilde{d}_\infty)}\}$ such that $\widetilde{d}_\infty(\mathcal{H}^n_{\sigma,R},\mathcal{H}^n_{\overline{\sigma},R})\le\varepsilon$. Then, by the definition of $\widetilde{d}_\infty$, we have $d_\infty(f,\mathcal{H}^n_{\overline{\sigma},R})\le\varepsilon$. And by the definition of $d_\infty$, there is $\overline{f}\in\mathcal{H}^n_{\overline{\sigma},R}$ satisfying $l_{\infty,n}(f,\overline{f})\le\varepsilon$.

By the definition of $\mathcal{N}(\varepsilon,\mathcal{H}^n_{\overline{\sigma},R},l_{\infty,n})$, there is $\widetilde{f}\in\{f^1_{\overline{\sigma}},\ldots,f^{\mathcal{N}(\varepsilon,\mathcal{H}^n_{\overline{\sigma},R},l_{\infty,n})}_{\overline{\sigma}}\}\subseteq\mathcal{H}^n_{\overline{\sigma},R}$ such that $l_{\infty,n}(\overline{f},\widetilde{f})\le\varepsilon$. Therefore, we can obtain

$$\mathcal{N}\left(2\varepsilon,\mathcal{H}^n_{\Gamma,R},l_{\infty,n}\right)\le\mathcal{N}_R\left(\varepsilon,\Gamma,\widetilde{d}_\infty\right)\cdot\sup_{\sigma\in\Gamma}\mathcal{N}\left(\varepsilon,\mathcal{H}^n_{\sigma,R},l_{\infty,n}\right). \tag{26}$$

Note that, for any $\sigma\in\Gamma$, there holds $\mathcal{N}(\varepsilon,\mathcal{H}^n_{\sigma,R},l_{\infty,n})\le(\mathcal{N}(\varepsilon,\mathcal{H}_{\sigma,R},l_{\infty,1}))^n$. Then the lemma follows. □

**Proposition 9.** *For $R=M/\sqrt{\lambda}$ and $\widehat{h}_i$, for $i=2,\ldots,n$, defined in* (10)*, there holds*

$$\begin{aligned}&P\left\{\frac{1}{n-1}\sum_{i=2}^n\left\{E^{\rho_i}\left(\widehat{h}_i\right)-\widehat{E}^{S_i}\left(\widehat{h}_i\right)\right\}\ge\varepsilon_1\right\}\\&\le\mathcal{N}\left(\frac{\varepsilon_1}{16(\kappa R+M)},\Gamma,d_\infty\right)\\&\quad\cdot\left(\sup_{\sigma\in\Gamma}\mathcal{N}\left(\frac{\varepsilon_1}{16(\kappa R+M)},\mathcal{H}_{\sigma,R},l_{\infty,1}\right)\right)^{n-1}\\&\quad\cdot\exp\left\{-\frac{(n-1)m\varepsilon_1^2}{8(\kappa R+M)^2}\right\}.\end{aligned} \tag{27}$$

*Proof.* By the definition of $\widehat{h}_i$, we have $\|\widehat{h}_i\|_{\widehat{\sigma}}\le R$, for $i=1,2,\ldots,n$. Then, there holds

$$\begin{aligned}&P\left\{\frac{1}{n-1}\sum_{i=2}^n\left\{E^{\rho_i}\left(\widehat{h}_i\right)-\widehat{E}^{S_i}\left(\widehat{h}_i\right)\right\}\ge\varepsilon_1\right\}\\&\le P\left\{\sup_{[f_2,\ldots,f_n]\in\mathcal{H}^{n-1}_{\Gamma,R}}\frac{1}{n-1}\sum_{i=2}^n\left\{E^{\rho_i}\left(f_i\right)-\widehat{E}^{S_i}\left(f_i\right)\right\}\ge\varepsilon_1\right\}.\end{aligned} \tag{28}$$

For $f=[f_2,\ldots,f_n]\in\mathcal{H}^{n-1}_{\Gamma,R}$, denote $L_{\mathbf{z}}(f)=(1/(n-1))\sum_{i=2}^n\{E^{\rho_i}(f)-\widehat{E}^{S_i}(f)\}$. Note that, for $f,g\in\mathcal{H}^{n-1}_{\Gamma,R}$, there holds

$$\begin{aligned}&\left|L_{\mathbf{z}}(f)-L_{\mathbf{z}}(g)\right|\\&=\frac{1}{n-1}\sum_{i=2}^n\left\{E^{\rho_i}(f)-\widehat{E}^{S_i}(f)\right\}\\&\quad-\frac{1}{n-1}\sum_{i=2}^n\left\{E^{\rho_i}(g)-\widehat{E}^{S_i}(g)\right\}\\&=\frac{1}{n-1}\sum_{i=2}^n\left\{\int_Z\left(f_i(x)-y\right)^2d\rho_i-\int_Z\left(g_i(x)-y\right)^2d\rho_i\right\}\\&\quad-\frac{1}{n-1}\sum_{i=2}^n\left\{\frac{1}{m}\sum_{j=1}^m\left(f_i\left(x^i_j\right)-y^i_j\right)^2\right.\\&\qquad\left.-\frac{1}{m}\sum_{j=1}^m\left(g_i\left(x^i_j\right)-y^i_j\right)^2\right\}\\&\le4(\kappa R+M)\cdot l_{\infty,n-1}(f,g).\end{aligned} \tag{29}$$

Therefore, we can find that $\mathcal{N}=\mathcal{N}(\varepsilon_1/8(\kappa R+M),\mathcal{H}^{n-1}_{\Gamma,R},l_{\infty,n-1})$ balls $\{\mathcal{H}^{n-1}_{k,R}\}_{k=1}^{\mathcal{N}}$ such that the center $\widetilde{f}^k$ of each ball and point $f^k$ in this ball satisfies $|L_{\mathbf{z}}(\widetilde{f}^k)-L_{\mathbf{z}}(f^k)|\le\varepsilon/2$. Therefore, the probability in (28) can be bounded by following expression:

$$\begin{aligned}&P\left\{\sup_{[f_2,\ldots,f_n]\in\bigcup_{k=1}^{\mathcal{N}(\varepsilon_1/8(\kappa R+M),\mathcal{H}^{n-1}_{\Gamma,R},l_{\infty,n})}\mathcal{H}^{n-1}_{k,R}}L_{\mathbf{z}}(f)\ge\varepsilon_1\right\}\\&\le\mathcal{N}\left(\frac{\varepsilon_1}{8(\kappa R+M)},\mathcal{H}^{n-1}_{\Gamma,R},l_{\infty,n}\right)\\&\quad\cdot\max_{k\in1,\ldots,\mathcal{N}}P\left\{\sup_{[f_2,\ldots,f_n]\in\mathcal{H}^{n-1}_{k,R}}L_{\mathbf{z}}(f)\ge\varepsilon_1\right\}\\&\le\mathcal{N}\left(\frac{\varepsilon_1}{8(\kappa R+M)},\mathcal{H}^{n-1}_{\Gamma,R},l_{\infty,n}\right)\\&\quad\cdot\max_{k\in1,\ldots,\mathcal{N}}P\left\{L_{\mathbf{z}}\left(\widetilde{f}^k\right)\ge\frac{\varepsilon_1}{2}\right\}.\end{aligned} \tag{30}$$

For the covering number $\mathcal{N}$, by Lemma 8, we have the estimate

$$\mathcal{N}\left(\frac{\varepsilon_1}{8(\kappa R+M)}, \mathcal{H}_{\Gamma,R}^{n-1}, l_{\infty,n}\right) \le \mathcal{N}\left(\frac{\varepsilon_1}{16(\kappa R+M)}, \Gamma, \widetilde{d}_\infty\right) \cdot \left(\sup_{\sigma\in\Gamma}\mathcal{N}\left(\frac{\varepsilon_1}{16(\kappa R+M)}, \mathcal{H}_{\sigma,R}, l_{\infty,1}\right)\right)^{n-1}. \tag{31}$$

Using Hoeffding inequality, for any fixed $\widetilde{f}^k$, we have

$$\begin{aligned} &P\left\{L_{\mathbf{z}}\left(\widetilde{f}^k\right) \ge \frac{\varepsilon_1}{2}\right\} \\ &= P\left\{\frac{1}{(n-1)m} \times \sum_{i=2}^{n}\sum_{j=1}^{m}\left\{\int_Z \left(\widetilde{f}_i^k(x)-y\right)^2 d\rho_i - \left(\widetilde{f}_i^k\left(x_j^i\right)-y_j^i\right)^2\right\} \ge \frac{\varepsilon_1}{2}\right\} \\ &\le \exp\left\{-\frac{(n-1)m\varepsilon_1^2}{8(\kappa R+M)^2}\right\}. \end{aligned} \tag{32}$$

Then, the proposition follows. □

Finally, we can obtain a bound for the last term of (16) by Proposition 9.

**Proposition 10.** *Let $\widehat{h}_i$, for $i = 2,\ldots,n$, be defined in* (10). *Assume, for all $\sigma \in \Gamma$ and for all $s > 0$, there holds*

$$\ln \mathcal{N}\left(\eta, \mathcal{H}_{\sigma,1}, l_{\infty,1}\right) \le C_0\left(\frac{1}{\eta}\right)^s. \tag{33}$$

*Then with confidence at least $1-\delta$, there holds*

$$\frac{1}{n-1}\sum_{i=2}^{n}\left\{E^{\rho_i}\left(\widehat{h}_i\right) - \widehat{E}^{S_i}\left(\widehat{h}_i\right)\right\} \le \varepsilon_0, \tag{34}$$

*where $\varepsilon_0$ is the solution of the following equation:*

$$\varepsilon_0^{2+s} - \frac{8(\ln(1/\delta))(\kappa R+M)^2}{(n-1)m}\varepsilon_0^s - \frac{8(\kappa R+M)^2}{m} \times C_0\left(16\left(\kappa R^2+MR\right)\right)^s - \frac{8(\kappa R+M)^2}{(n-1)m}\ln \mathcal{N}_R\left(\frac{\varepsilon_0}{16(\kappa R+M)}, \Gamma, d_\infty\right) = 0. \tag{35}$$

*Proof.* Let

$$\delta = \mathcal{N}_R\left(\frac{\varepsilon_1}{16(\kappa R+M)}, \Gamma, d_\infty\right) \cdot \left(\sup_{\sigma\in\Gamma}\mathcal{N}\left(\frac{\varepsilon_1}{16(\kappa R+M)}, \mathcal{H}_{\sigma,R}, l_\infty^1\right)\right)^{n-1} \cdot \exp\left\{-\frac{(n-1)m\varepsilon_1^2}{8(\kappa R+M)^2}\right\}. \tag{36}$$

By condition on $\ln \mathcal{N}(\eta, \mathcal{H}_{\sigma,1}, l_{\infty,1}) \le C_0(1/\eta)^s$, we have

$$\ln \mathcal{N}\left(\frac{\varepsilon_1}{16(\kappa R+M)}, \mathcal{H}_{\sigma,R}, l_{\infty,1}\right) \le C_0\left(\frac{16(\kappa R^2+MR)}{\varepsilon_1}\right)^s. \tag{37}$$

Then, $\varepsilon_1$ is not larger than $\varepsilon_0$ in the following equation:

$$\ln\delta = \ln \mathcal{N}_R\left(\frac{\varepsilon_0}{16(\kappa R+M)}, \Gamma, d_\infty\right) + (n-1)C_0\left(\frac{16(\kappa R^2+MR)}{\varepsilon_0}\right)^s - \frac{(n-1)m}{8(\kappa R+M)^2}\varepsilon_0^2. \tag{38}$$

Then by Proposition 9, the proposition follows. □

*Remark 11.* Compare multitask learning with multiple Hilbert spaces to single task learning with multiple Hilbert space.

Recall that the least square regularized regression in $\{\mathcal{H}_\sigma, \sigma \in \Gamma\}$ for single task is defined as

$$\left(\widehat{\sigma}_1, \widehat{h}_1\right) = \arg\min_{h_1\in\mathcal{H}_\sigma,\sigma\in\Gamma}\left\{\widehat{E}^{S_1}(h_1) + \lambda\|h_1\|_\sigma^2\right\}. \tag{39}$$

$E^{\rho_1}(\widehat{h}_1) - E^{\rho_1}(f_{\rho_1})$ can be bounded by the sum of regularization error and sample error with similar method in Proposition 6. In the sample error, the term most difficult to estimated will be $E^{\widehat{S}_1}(\widehat{h}_1) - E^{\rho_1}(\widehat{h}_1)$, because function $\widehat{h}_1$ changed with the $S_1$ runs over function set $\{\mathcal{H}_\sigma, \sigma \in \Gamma\}$. By the same method in Proposition 10, with confidence $1-\delta$, this term can be bounded by the solution $\widetilde{\varepsilon}_0$ of the following equation:

$$\widetilde{\varepsilon}_0^{2+s} - \frac{8(\ln(1/\delta))(\kappa R+M)^2}{m}\widetilde{\varepsilon}_0^s - \frac{8(\kappa R+M)^2}{m} \times C_0\left(16\left(\kappa R^2+MR\right)\right)^s - \frac{8(\kappa R+M)^2}{m}\ln \mathcal{N}_R\left(\frac{\widetilde{\varepsilon}_0}{16(\kappa R+M)}, \Gamma, d_\infty\right) = 0. \tag{40}$$

Obviously, we have $\varepsilon_0 < \widetilde{\varepsilon}_0$. Therefore, multitask learning algorithm has potential faster learning rate.

*Remark 12.* Comparing multitask learning with multiple Hilbert spaces to single task learning with single Hilbert space.

In this paper, we set the hypothesis space as a set of Hilbert spaces. It is well known that hypothesis space with more

functions has stronger approximation ability and bigger complexity. Therefore, the regularization error may be smaller and sample error may be larger than that of algorithms with a single Hilbert space being the hypothesis space.

For least square regularized regression with a single Hilbert space $\mathcal{H}_\sigma$ for single task, with confidence $1-\delta$, the largest term in sample error can be bounded by the solution $\bar{\varepsilon}_0$ of the following equation:

$$\bar{\varepsilon}_0^{2+s} - \frac{8\left(\ln\left(1/\delta\right)\right)\left(\kappa R + M\right)^2}{m}\bar{\varepsilon}_0^{s} - \frac{8(\kappa R + M)^2}{m}C_0\left(16(\kappa R^2 + MR)\right)^{s} = 0. \tag{41}$$

If we assume $n = \mathcal{O}(m^{\zeta})$ with some $\zeta > 0$ large enough, $(8(\kappa R + M)^2/(n-1)m)\ln \mathcal{N}_R(\varepsilon_0/16(\kappa R + M), \Gamma, d_\infty)$ in Proposition 10 can converge to 0 fast. Then, we can obtain $\varepsilon_0 \approx \bar{\varepsilon}_0$, while the regularization error of multitask learning with multiple Hilbert spaces is smaller. Trading off the regularization error and sample error, we can obtain potential faster learning rate than that of single task learning.

## Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant ZY1118, National Technology Support Program under Grant 2012BAH05B01, and the National Science Foundation of China under Grants 11101024 and 11171014.

## References

[1] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[2] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI '10)*, pp. 733–742, Catalina Island, Calif, USA, July 2010.

[3] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, New York, NY, USA, August 2004.

[4] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[5] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient L2,1-norm minimization," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI '09)*, pp. 339–348, Montreal, Canada, June 2009.

[6] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, no. 1, pp. 83–99, 2004.

[7] A. Argyriou, C. A. Micchelli, M. Pontil, and Y. M. Ying, "A spectral regularization framework for multi-task structure learning," in *Proceedings of the 21st Annual Conference on Advances in Neural Information Processing Systems (NIPS '07)*, December 2007.

[8] J. Guinney, Q. Wu, and S. Mukherjee, "Estimating variable structure and dependence in multitask learning via gradients," *Machine Learning*, vol. 83, no. 3, pp. 265–287, 2011.

[9] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, "Multi-task learning for classification with Dirichlet process priors," *Journal of Machine Learning Research*, vol. 8, pp. 35–63, 2007.

[10] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[11] J. Baxter, "Learning internal representations," in *Proceedings of the Workshop on Computational Learning Theory (COLT '95)*, Morgan Kaufmann, San Mateo, Calif, USA, 1995.

[12] N. Intrator and S. Edelman, "Making a low-dimensional representation suitable for diverse tasks," *Connection Science*, vol. 8, no. 2, pp. 205–224, 1996.

[13] S. Thrun, "Is learning the n-th thing any easier than learning the first?" in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '96)*, D. Touretzky and M. Mozer, Eds., 1996.

[14] T. Heskes, "Solving a huge number of similar tasks: a combination of multi-task learning and a hierarchical Bayesian approach," in *Proceedings of the International Conference on Machine Learning (ICML '98)*, 1998.

[15] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, "Exploiting unrelated tasks in multi-task learning," in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS '12)*, pp. 951–959, La Palma, Spain, 2012.

[16] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[17] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, pp. 1817–1853, 2005.

[18] S. Ben-David and R. S. Borbely, "A notion of task relatedness yielding provable multiple-task learning guarantees," *Machine Learning*, vol. 73, no. 3, pp. 273–287, 2008.

[19] M. Solnon, S. Arlot, and F. Bach, "Multi-task regression using minimal penalties," *Journal of Machine Learning Research*, vol. 13, pp. 2773–2812, 2012.

[20] Y. M. Ying and D.-X. Zhou, "Learnability of Gaussians with flexible variances," *Journal of Machine Learning Research*, vol. 8, pp. 249–276, 2007.

[21] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 171–192, 2006.