Appl. Comput. Harmon. Anal. ••• (••••) •••–•••

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Distributed learning with multi-penalty regularization ☆

Zheng-Chu Guo [a], Shao-Bo Lin [b,*], Lei Shi [c]

[a] *School of Mathematical Sciences, Zhejiang University, Hangzhou, 310027, China*
[b] *Department of Statistics, Wenzhou University, Wenzhou, 325035, China*
[c] *Shanghai Key Laboratory for Contemporary Applied Mathematics, School of Mathematical Sciences, Fudan University, Shanghai, 200433, China*

ARTICLE INFO

Article history:
Received 29 November 2016
Received in revised form 1 June 2017
Accepted 10 June 2017
Available online xxxx
Communicated by Ding-Xuan Zhou

Keywords:
Learning theory
Multi-penalty regularization
Manifold regularization
Integral operator

ABSTRACT

In this paper, we study distributed learning with multi-penalty regularization based on a divide-and-conquer approach. Using Neumann expansion and a second order decomposition on difference of operator inverses approach, we derive optimal learning rates for distributed multi-penalty regularization in expectation. As a byproduct, we also deduce optimal learning rates for multi-penalty regularization, which was not given in the literature. These results are applied to the distributed manifold regularization and optimal learning rates are given.

1. Introduction

In recent years, we have witnessed fast growing interest in studying multi-penalty regularization such as the elastic-net regularization [35] and manifold regularization [4] in machine learning. Compared with one-penalty regularization like the regularized least squares [10] and kernel LASSO [25], multi-penalty regularization is recognized to be capable of exploring more delicate information from data, since it incorporates different priori knowledge into different penalties. With this, multi-penalty regularization was widely used in the image reconstruction [22], Earth gravity potential reconstruction [33] and option pricing [9].

However, apart from its promising applications, the theoretical feasibility of multi-penalty regularization is questionable since optimal learning rates are only achieved by some one-penalty regularization algorithms, to just name a few, [2,6,8,32,30,16,17,15]. Furthermore, as there are more than one parameters, the compu-

☆ The research was supported by the National Natural Science Foundation of China [Grant Nos. 11401524, 11531013, 61502342, 11571078, 11631015]. Lei Shi is also supported by the Joint Research Fund by National Natural Science Foundation of China and Research Grants Council of Hong Kong (Project No. 11461161006 and Project No. CityU 104012) and Zhuo Xue program of Fudan University. Part of the work was carried out while Zheng-Chu Guo was visiting Shanghai Key Laboratory for Contemporary Applied Mathematics. Authors contributed equally to this paper and are listed alphabetically.

* Corresponding author.
E-mail address: sblin1983@gmail.com (S.-B. Lin).

http://dx.doi.org/10.1016/j.acha.2017.06.001
1063-5203/

tational burden to get an estimator based on multi-penalty regularization is very high, especially when the size of data is huge. This motivates us to consider the distributed multi-penalty regularization algorithm based on a divide-and-conquer strategy [34] to reduce the computational burden.

Let $\mathcal{H}_K$ be a reproducing kernel Hilbert space (RKHS) induced by a Mercer kernel K on a compact metric space $\mathcal{X}$. Given a sample $D = \{(x_i, y_i)\}_{i=1}^{|D|} \subset \mathcal{X} \times \mathcal{Y}$ with output space $\mathcal{Y} = \mathbb{R}$, we consider the kernel-based multi-penalty regularized least squares (MP-RLS):

$$f_{D,\lambda_1,\lambda_2} := \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x,y)\in D} (f(x) - y)^2 + \lambda_1 \|f\|_K^2 + \lambda_2 \|J_D f\|_K^2 \right\}, \tag{1.1}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are regularization parameters, J_D is a linear operator on $\mathcal{H}_K$ that may depend on the data D, $\|\cdot\|_K$ is the norm of $\mathcal{H}_K$ and $|D|$ denotes the cardinality of D. When $\lambda_2 = 0$, algorithm (1.1) coincides with the regularized least squares (RLS) [10]. If J_D is a bounded operator depending on the graphic Laplacian [4], algorithm (1.1) is the manifold regularization.

Similar as the distributed regularized least squares [34,17], distributed MP-RLS begins with partitioning the data set D into m disjoint subsets $\{D_j\}_{j=1}^m$. Then it assigns each data subset D_j to one machine (called local machine) to produce a local estimator $f_{D_j,\lambda_1,\lambda_2}$ using algorithm (1.1). Finally, these local estimators are transmitted to a central machine, and a global estimator $\overline{f}_{D,\lambda_1,\lambda_2}$ is synthesized by taking weighted averaging

$$\overline{f}_{D,\lambda_1,\lambda_2} = \sum_{j=1}^{m} \frac{|D_j|}{|D|} f_{D_j,\lambda_1,\lambda_2}. \tag{1.2}$$

Since local estimators can be obtained by parallel computation and the size of data in each local machine is much smaller than $|D|$ when m is large, the computational burden of algorithm (1.1) is essentially reduced via the proposed distributed learning strategy (1.2).

The key difficulty in our analysis is that algorithm (1.2) involves an extra operator J_D, which makes the existing theoretical analysis [6,17] unavailable. Utilizing the Neumann expansion and the second order decomposition on the difference of operator inverses [17], we present a novel error analysis to quantify the performance of MP-RLS. In our analysis, except for the boundedness, we do not impose any other conditions on the operator J_D. This approach establishes a unified analysis framework for MP-RLS with a wide range of J_D and succeeds in deducing optimal learning rates for algorithm (1.2). As a by-product, we deduce optimal learning rates for MP-RLS, which improves the learning rates in [1] from sub-optimal to optimal. As an application of our theory, we propose a novel distributed manifold regularization algorithm and derive its optimal learning rate. We find that the unlabeled data in manifold regularization [4] not only plays a crucial role in embodying the manifold feature of the input space, but also benefits in enlarging the number of local machines in distributed learning and relaxing the restriction to the regression function.

2. Optimal learning rates for distributed MP-RLS

In learning theory [8], $D = \{x_i, y_i\}_{i=1}^{|D|}$ are drawn independently according to ρ, a Borel probability measure on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The objective of learning is to construct an estimator f_D based on D to approximate the regression function defined by

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \qquad x \in \mathcal{X},$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

where $\rho(y|x)$ denotes the conditional distribution at x induced by ρ. Let ρ_X be the marginal distribution of ρ on $\mathcal{X}$ and $(L^2_{\rho_X}, \|\cdot\|_\rho)$ be the Hilbert space of ρ_X square integrable functions on $\mathcal{X}$. Our aim is to bound $\|\overline{f}_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho$.

2.1. Assumptions and main results

To derive a concrete learning rate, some assumptions concerning the regularity of f_ρ, capacity of $\mathcal{H}_K$, type of noise and property of J_D should be imposed. Define the integral operator L_K on $\mathcal{H}_K$ or $L^2_{\rho_X}$ associated with the Mercer kernel K by

$$L_K(f) = \int_X f(x) K_x d\rho_X,$$

where $K_x(\cdot) := K(x, \cdot)$. Our first assumption describes the regularity property of the regression function.

Assumption 1. For some $r > 0$, there holds

$$f_\rho = L_K^r(h_\rho) \quad \text{with } h_\rho \in L^2_{\rho_X}, \tag{2.1}$$

where L_K^r denotes the r-th power of L_K on $L^2_{\rho_X}$.

Since $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ is a compact and positive operator, L_K^r is well defined. Assumption 1 is a widely used source condition which has been adopted in [29,2,6,7,26,17,15] to quantify learning rates for regularization algorithms. It should be noted that (2.1) with $r = \frac{1}{2}$ is equivalent to $f_\rho \in \mathcal{H}_K$.

Our second assumption concerns the property of the noise.

Assumption 2. If (2.1) holds with $r \geq \frac{1}{2}$, then $\int_{\mathcal{Y}} y^2 d\rho < \infty$ and

$$\int_{\mathcal{Y}} \left(e^{\frac{|y - f_\rho(x)|}{M}} - \frac{|y - f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\gamma^2}{2M^2}, \qquad \forall x \in \mathcal{X}. \tag{2.2}$$

Otherwise, $|y| \leq M$ almost surely, where M and γ are positive constants.

It can be found in [2] that (2.2) implies the so-called Bernstein condition

$$\int_{\mathcal{Y}} (y - f_\rho(x))^\ell d\rho(y|x) \leq \frac{1}{2} \ell! \gamma^2 M^{\ell-2}, \qquad \forall \ell \geq 2, x \in \mathcal{X},$$

which is a broad model for the noise of the output y, containing the uniformly bounded, Gaussian or sub-Gaussian noise [6,3,14].

We use the effective dimension $\mathcal{N}(\lambda_1)$ to measure the capacity of $\mathcal{H}_K$ with respect to ρ_X, which is defined to be the trace of the operator $(\lambda_1 I + L_K)^{-1} L_K$, that is

$$\mathcal{N}(\lambda_1) = \mathrm{Tr}((L_K + \lambda_1 I)^{-1} L_K), \qquad \lambda_1 > 0.$$

Our third assumption focuses on the rate of decaying of $\mathcal{N}(\lambda_1)$.

Assumption 3. There exists a constant $C_0 > 0$ such that for all $\lambda_1 > 0$,

$$\mathcal{N}(\lambda_1) \leq C_0 \lambda_1^{-\beta}, \quad \text{for some} \quad 0 < \beta \leq 1. \tag{2.3}$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

It is obvious that (2.3) always holds with $\beta = 1$ and $C_0 = \mathrm{Tr}(L_K) \le \kappa^2$. For $0 < \beta < 1$, let $\{(\sigma_\ell, \phi_\ell)\}_\ell$ be a set of normalized eigenpairs of L_K on $\mathcal{H}_K$ with $\{\phi_\ell\}_{\ell=1}^\infty$ forming an orthonormal basis of $\mathcal{H}_K$, and let

$$L_K = \sum_{\ell=1}^{\infty} \sigma_\ell \langle \cdot, \phi_\ell \rangle_K \phi_\ell$$

be the spectral decomposition. It can be found in [15] that $\sigma_n \le C_0 n^{-1/\beta}$ for some $0 < \beta < 1$ implies (2.3). Thus, the effective dimension decaying assumption is weaker than the widely used eigenvalue decaying assumption [6,30].

To derive the learning rate, we also need some restrictions to the linear operator J_D.

Assumption 4. There is a constant $c_J > 0$ independent of $|D|$ such that $\|J_D^T J_D\| \le c_J$ holds almost surely, where J_D^T denotes the adjoint operator of J_D and $\|\cdot\|$ is the operator norm.

Assumption 4 only concerns the boundedness of J_D, which makes our analysis more general than the existing analysis concerning multi-penalty regularization [12,24,11]. In the next section, we will show that the manifold regularization is a special example of algorithm (1.1).

Based on these assumptions, we can derive the following optimal learning rate for algorithm (1.2).

Theorem 2.1. *Under Assumptions 1–4 with* $\frac{1}{2} \le r \le 1$ *and* $0 < \beta \le 1$, *if* $|D_1| = |D_2| = \cdots = |D_m|$, $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$, $\lambda_2 = \frac{1}{2c_J}|D|^{-\frac{2r}{2r+\beta}}$ *and*

$$m \le |D|^{\frac{2r-1}{2r+\beta}}, \tag{2.4}$$

then

$$E\left[\|\overline{f}_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] \le \hat{C}|D|^{-\frac{2r}{2r+\beta}},$$

where $\hat{C}$ *is a constant independent of* m *or* $|D|$ *and will be given in the proof explicitly.*

Theorem 2.1 together with [6, Theorem 2] shows that the distributed MP-RLS algorithm (1.2) can reach the optimal learning rate, provided m satisfies (2.4). As a by-product, we can deduce the optimal learning rate for algorithm (1.1).

Theorem 2.2. *Let* $0 < \delta < 1$. *Under Assumptions 1–4 with* $0 < r \le 1$ *and* $0 < \beta \le 1$, *if* $\lambda_1 = |D|^{-\frac{1}{\max\{2r,1\}+\beta}}$ *and* $\lambda_2 = \frac{1}{2c_J}|D|^{-\frac{\max\{2r,1\}}{\max\{2r,1\}+\beta}}$, *then with confidence at least* $1-\delta$, *there holds*

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \le \tilde{C}|D|^{-\frac{r}{\max\{2r,1\}+\beta}} \left(\log\frac{6}{\delta}\right)^3, \tag{2.5}$$

where $\tilde{C}$ *is a positive constant independent of* $|D|$ *or* δ. *Moreover*

$$E\left[\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] = \mathcal{O}\left(|D|^{-\frac{2r}{\max\{2r,1\}+\beta}}\right). \tag{2.6}$$

When $\frac{1}{2} \le r \le 1$, the learning rate exhibited in Theorem 2.2 is optimal. However, when $0 < r < \frac{1}{2}$, i.e. $f_\rho \notin \mathcal{H}_K$, our result is sub-optimal.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

2.2. Related work and discussions

Distributed and parallel computation is a hot topic in the coming big data era. Theoretical analysis for specified distributed learning algorithms has triggered enormous research activities in the statistical and machine learning communities. In particular, using a relation between the stability and generalization, [23] deduced a variance estimate for distributed conditional maximum entropy models. Utilizing a matrix decomposition approach, [34] derived optimal learning rates for distributed RLS under some eigenfunction assumptions, which were removed in [17] by employing a novel integral operator approach. Motivated by [17], properties of various integral operators were revealed and applied to derive optimal learning rates for distributed spectral algorithms [15] and distributed gradient descent algorithms [18]. In the present paper, we also adopt the integral operator approach to study optimal learning rates for algorithm (1.2). In particular, under Assumptions 1 to 3 with $\frac{1}{2} \leq r \leq 1$ and $0 < \beta \leq 1$, we showed that distributed learning with MP-RLS can reach the optimal learning of order $\mathcal{O}(|D|^{-2r/(2r+\beta)})$, provided that J_D is a bounded linear operator, m satisfies (2.4), $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$ and $\lambda_2 = \frac{1}{c_J}|D|^{-\frac{2r}{2r+\beta}}$. Since λ_2 is comparable with λ_1 for small r, two penalties in algorithm (1.1) play comparable roles in our analysis.

Studying the convergence rate of multi-penalty regularization also attracted growing attention in the inverse problems community. Optimal convergence rates for various multi-penalty regularization were derived in [19,20,31,12,24]. As an ill-posed problem solver, multi-penalty regularization can conquer the saturation phenomenon of the Tikhonov regularization [12], can produce sparse solutions [31] and is more robust than one-penalty regularization [19]. In the interesting work [21], some operator theory for multi-penalty regularization was extended to learning theory. Motivated by [21], the recent paper [1] presented error estimates for algorithm (1.1) under Assumptions 1–4 with $\frac{1}{2} \leq r \leq 1$ and $\beta = 1$, saying that there exists a constant C' independent of $|D|$, δ, λ_1 or λ_2 such that

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \leq C'|D|^{-r/(2r+1)} \log \frac{4}{\delta} \tag{2.7}$$

holds with confidence at least $1-\delta$ for some parameters λ_1 and λ_2. Since $\frac{1}{2} \leq r \leq 1$, the above learning rate is slower than $\mathcal{O}(|D|^{-1/3})$. Compared with (2.7), Theorem 2.2 provides learning rates for MP-RLS for $0 < \beta \leq 1$. It can be found in (2.5) that smaller β implies faster learning rate.

In this paper, we only impose the boundedness assumption to the linear operator J_D. It is interesting to derive learning rate for MP-RLS with J_D satisfying other assumptions. An interesting attempt is to borrow the idea from [12,24] to utilize multi-penalty regularization to circumvent the saturation for RLS in the realm of machine learning.

3. Applications to distributed manifold regularization

In manifold regularization [4], we are given both labeled data D and unlabeled data $\tilde{D}(x) = \{\tilde{x}_1, \cdots, \tilde{x}_{|\tilde{D}|}\}$ which are drawn independently according to ρ_X. If the support of the marginal distribution ρ_X is a compact sub-manifold $\mathcal{M} \subset \mathcal{X}$, the aim is to add another penalty depending on $\tilde{D}(x)$ to RLS to incorporate the manifold feature. Let

$$D^* := D \cup \tilde{D} = \{x_i^*, y_i^*\}_{i=1}^{|D^*|},$$

where $\tilde{D} = \{(\tilde{x}_i, 0)\}_{\tilde{x}_i \in \tilde{D}(x)}$. Denote $S_D : \mathcal{H}_K \to \mathbb{R}^{|D|}$ (or $L^2_{\rho_X} \to \mathbb{R}^{|D|}$) by

$$S_D f := (f(x_i))_{(x_i,y_i) \in D}$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

and $S_D^T : \mathbb{R}^{|D|} \to \mathcal{H}_K$ (or $\mathbb{R}^{|D|} \to L^2_{\rho_X}$) by

$$S_D^T \mathbf{c} := \frac{1}{|D|} \sum_{(x_i,y_i)\in D} c_i K_{x_i}, \ \mathbf{c} = (c_1, \dots, c_{|D|}) \in \mathbb{R}^{|D|}.$$

Define further $W_{D^*} = (w_{i,j})_{i,j=1}^{|D^*|}$ as the matrix whose elements are the edge weights in the data adjacency graph [4, p. 2405] and $V_{D^*} = (v_{i,j})_{i,j=1}^{|D^*|}$ as the diagonal matrix with $v_{i,i} = \sum_{j=1}^{|D^*|} w_{i,j}$. Define the graphic Laplace as $L_{D^*} = V_{D^*} - W_{D^*}$. The following manifold regularization is defined in [4, Eqs. (4)] by

$$\begin{aligned} f^{mr}_{D^*,\lambda_1,\lambda_2} &:= \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x_i,y_i)\in D} (f(x_i) - y_i)^2 + \lambda_1 \|f\|_K^2 + \frac{\lambda_2}{|D^*|^2} \sum_{i,j=1}^{|D^*|} (f(x_i) - f(x_j))^2 w_{i,j} \right\} \\ &= \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x_i,y_i)\in D} (f(x_i) - y_i)^2 + \lambda_1 \|f\|_K^2 + \frac{\lambda_2}{|D^*|} \|(S_{D^*}^T L_{D^*} S_{D^*})^{\frac{1}{2}} f\|_K^2 \right\}. \end{aligned} \tag{3.1}$$

Setting $J_{D^*} = (S_{D^*}^T L_{D^*} S_{D^*})^{\frac{1}{2}} / \sqrt{|D^*|}$, it is easy to derive (similar as [1, Prop. 3.4], but with a slight change) that

$$\|J_{D^*}^T J_{D^*}\| \le 2w\kappa^2 \tag{3.2}$$

holds almost surely, where $w := \max_{1\le i,j\le |D^*|} w_{i,j}$ and $\kappa := \max\{1, \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}\}$. This implies that manifold regularization (3.1) is a special example of algorithm (1.1). Then the optimal learning rates of manifold regularization and its corresponding distributed version follow from Theorem 2.2 and Theorem 2.1.

In (3.1), unlabeled data $\tilde{D}$ are only involved in the penalty. The recent study in [7,5,18] showed that unlabeled data play a crucial role in distributed learning in terms of enlarging the range of m and f_ρ. Hence, we present a slight modification of the original manifold regularization algorithm (3.1) by involving the unlabeled data in the least squares term. For each $j = 1, \cdots, m$, we set

$$D_j^* := D_j \cup \tilde{D}_j = \{x_i^*, y_i^*\}_{i=1}^{|D_j^*|}$$

with

$$x_i^* := \begin{cases} x_i, & \text{if } (x_i, y_i) \in D_j, \\ \tilde{x}_i, & \text{if } \tilde{x}_i \in \tilde{D}_j(x), \end{cases} \quad \text{and} \quad y_i^* := \begin{cases} \frac{|D_j^*|}{|D_j|} y_i, & \text{if } (x_i, y_i) \in D_j, \\ 0, & \text{otherwise,} \end{cases}$$

and $D^* = \bigcup_{j=1}^m D_j^*$. The modified distributed manifold regularization algorithm can be stated as

$$\overline{f}^*_{D^*,\lambda_1,\lambda_2} := \sum_{j=1}^m \frac{|D_j^*|}{|D^*|} f^*_{D_j^*\lambda_1,\lambda_2} \tag{3.3}$$

with

$$f^*_{D_j^*,\lambda_1,\lambda_2} := \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D_j^*|} \sum_{(x_i,y_i)\in D_j^*} (f(x_i) - y_i)^2 + \lambda_1 \|f\|_K^2 + \lambda_2 \|J_{D_j^*} f\|_K^2 \right\}. \tag{3.4}$$

The following theorem is the main result of this section.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Theorem 3.1. *Under Assumptions 1–3 with* $0 < r \leq 1$, $0 < \beta \leq 1$ *and* $r + \beta \geq \frac{1}{2}$, *if* $|D^*| \geq |D|^{\frac{1+\beta}{2r+\beta}}$, $|D_1| = |D_2| = \cdots = |D_m|$, $|D_1^*| = |D_2^*| = \cdots = |D_m^*|$, $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$, $\lambda_2 = \frac{1}{4\omega\kappa^2}|D|^{-\frac{\max\{2r,1\}}{2r+\beta}}$, *and* m *satisfies*

$$m \leq \min\left\{|D^*|\,|D|^{-\frac{1+\beta}{2r+\beta}}, |D|^{\frac{2r+2\beta-1}{2r+\beta}}\right\},$$

then

$$E\left[\|\overline{f}^*_{D^*,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] = \mathcal{O}\left(|D|^{-\frac{2r}{2r+\beta}}\right). \tag{3.5}$$

We see from Theorem 2.1 that the restriction on the number of local machines is a bit strict, since that when $r = \frac{1}{2}$, to achieve the optimal learning rate, the number of local machines is a constant. However, in Theorem 3.1, utilizing the unlabeled data in manifold regularization, we show that this restriction can be relaxed. In fact, if $|D^*| = |D|$, i.e., we consider distributed learning with the classical manifold regularization (3.1), the number of local machines is at most $|D|^{\frac{2r-1}{2r+\beta}}$, which is exactly the same as that of Theorem 2.1. If we employ the unlabeled data in the least square term in algorithm (3.3) and let $|D^*| \geq |D|^{\frac{2r+3\beta}{2r+\beta}}$, then the number of local machines is less than

$$\min\left\{|D^*|\,|D|^{-\frac{1+\beta}{2r+\beta}}, |D|^{\frac{2r+2\beta-1}{2r+\beta}}\right\} = |D|^{\frac{2r+2\beta-1}{2r+\beta}},$$

which is essentially larger than $|D|^{\frac{2r-1}{2r+\beta}}$.

When $r \geq 1/2$, the condition $|D^*| \geq |D|^{\frac{1+\beta}{2r+\beta}}$ and $r + \beta \geq 1/2$ always hold. If $0 < r < 1/2$ and $m = 1$, (3.5) shows that the learning rate of algorithm (1.1) can be improved from $\mathcal{O}\left(|D|^{-\frac{2r}{1+\beta}}\right)$ to $\mathcal{O}\left(|D|^{-\frac{2r}{2r+\beta}}\right)$, which shows the power of unlabeled data in our modification of manifold regularization.

4. Operator product norm estimates and error decomposition

We use the integral operator approach [27–29] to analyze the learning ability of algorithms (1.1) and (1.2). Let $L_{K,D}$ be the data-dependent approximation of L_K defined by

$$L_{K,D}f := S_D^T S_D f = \frac{1}{|D|}\sum_{(x,y)\in D} f(x)K_x.$$

Then it is easy to check [1] that the solution of algorithm (1.1) is given by

$$f_{D,\lambda_1,\lambda_2} = \left(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D\right)^{-1} S_D^T y_D. \tag{4.1}$$

4.1. Operator product norm estimates

The main novelty in our analysis is to use the operator product norm to quantify the difference between the solutions of MP-RLS and RLS, i.e., to quantify the similarity between $\left(L_K + \lambda_1 I + \lambda_2 J_D^T J_D\right)^{-1}$ and $(L_K + \lambda_1 I)^{-1}$, and use it to bound the operator product norm of $\|(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)\|$. Based on the Neumann expansion, we obtain the following proposition.

Proposition 4.1. *Under Assumption 4, if* $\lambda_1, \lambda_2 \in (0,1)$ *satisfy* $2c_J\lambda_2 = \lambda_1^{\max\{1,2r\}}$ *for some* $r \in (0,1]$, *then there holds almost surely*

$$\|(L_K + \lambda_1 I)(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\| \leq 2. \tag{4.2}$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Proof. Denote $A = L_K + \lambda_1 I$ and $B = L_K + \lambda_1 I + \lambda_2 J_D^T J_D$. Since $2c_J\lambda_2 = \lambda_1^{\max\{1,2r\}}$ and $0 < \lambda_1 < 1$, we have almost surely

$$\|A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}\| = \left\|(L_K+\lambda_1 I)^{-\frac{1}{2}}(-\lambda_2 J_D^T J_D)(L_K+\lambda_1 I)^{-\frac{1}{2}}\right\|$$
$$\leq \lambda_1^{-\frac{1}{2}}\lambda_2\|J_D^T J_D\|\lambda_1^{-\frac{1}{2}} \leq c_J\lambda_2\lambda_1^{-1} \leq \frac{1}{2}.$$

Then, $(I - A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}})^{-1}$ is well defined and

$$AB^{-1} = A^{\frac{1}{2}}(I - A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}})^{-1}A^{-\frac{1}{2}}.$$

The Neumann expansion implies that

$$A^{\frac{1}{2}}(I - A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}})^{-1}A^{-\frac{1}{2}} = A^{\frac{1}{2}}\sum_{n=0}^{\infty}[A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}]^n A^{-\frac{1}{2}}$$
$$= I + (A-B)A^{-1} + A^{\frac{1}{2}}\sum_{n=2}^{\infty}[A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}]^n A^{-\frac{1}{2}}$$
$$= I + (A-B)A^{-1} + (A-B)A^{-\frac{1}{2}}\sum_{n=0}^{\infty}[A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}]^n A^{-\frac{1}{2}}(A-B)A^{-1}.$$

But

$$\|(A-B)A^{-1}\| = \|(-\lambda_2 J_D^T J_D)(L_K+\lambda_1 I)^{-1}\| \leq c_J\lambda_2\lambda_1^{-1} \leq \frac{1}{2},$$

$$\|(A-B)A^{-\frac{1}{2}}\| = \|(-\lambda_2 J_D^T J_D)(L_K+\lambda_1 I)^{-\frac{1}{2}}\| \leq c_J\lambda_2\lambda_1^{-1/2} \leq \frac{1}{2}\sqrt{\lambda_1},$$

$$\|A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}\| = \|(L_K+\lambda I)^{-\frac{1}{2}}(-\lambda_2 J_D^T J_D)(L_K+\lambda I)^{-\frac{1}{2}}\| \leq c_J\lambda_2\lambda_1^{-1} \leq \frac{1}{2}$$

and

$$\|A^{-\frac{1}{2}}(A-B)A^{-1}\| = \|A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}}\| \leq \frac{1}{2\sqrt{\lambda_1}}.$$

We obtain

$$\|AB^{-1}\| \leq 1 + \|(A-B)A^{-1}\|$$
$$+ \|(A-B)A^{-\frac{1}{2}}\|\sum_{n=0}^{\infty}\|[A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}]\|^n\|A^{-\frac{1}{2}}(A-B)A^{-1}\|$$
$$\leq 1 + \frac{1}{2} + \frac{1}{2}\sqrt{\lambda_1}\times\frac{1}{1-\frac{1}{2}}\times\frac{1}{2\sqrt{\lambda_1}} = 2,$$

which completes the proof of Proposition 4.1. □

Let

$$\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D} := \left\|(L_K+\lambda_1 I+\lambda_2 J_D^T J_D)(L_{K,D}+\lambda_1 I+\lambda_2 J_D^T J_D)^{-1}\right\|.$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

The main tools to bound $\mathcal{Q}_{D,\lambda,J}$ are the second order decomposition for the difference of operator inverses presented in [17,15], Proposition 4.1 and the following Lemma 4.2, which was proved in [17].

Lemma 4.2. *Let $0 < \delta < 1$ and D be a sample drawn independently according to ρ. With confidence at least $1-\delta$, there holds*

$$\left\| (L_K + \lambda_1 I)^{-\frac{1}{2}} \{L_K - L_{K,D}\} \right\| \leq \mathcal{B}_{|D|,\lambda_1} \log(2/\delta), \tag{4.3}$$

where

$$\mathcal{B}_{|D|,\lambda_1} := \frac{2\kappa}{\sqrt{|D|}} \left\{ \frac{\kappa}{\sqrt{|D|\lambda_1}} + \sqrt{\mathcal{N}(\lambda_1)} \right\}.$$

If A and B are invertible operators on a Banach space, the second order decomposition proposed in [17] shows

$$\begin{aligned} A^{-1} - B^{-1} &= B^{-1}(B-A)A^{-1}(B-A)B^{-1} + B^{-1}(B-A)B^{-1} \\ &= B^{-1}(B-A)B^{-1}(B-A)A^{-1} + B^{-1}(B-A)B^{-1}. \end{aligned} \tag{4.4}$$

This implies the following decomposition of the operator product

$$BA^{-1} = (B-A)B^{-1}(B-A)A^{-1} + (B-A)B^{-1} + I. \tag{4.5}$$

Inserting $A = L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D$ and $B = L_K + \lambda_1 I + \lambda_2 J_D^T J_D$ to (4.5), we obtain the following upper bound for $\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D}$.

Proposition 4.3. *Let $0 < \delta < 1$. Under Assumption 4, if $\lambda_1, \lambda_2 \in (0,1)$ satisfy $2c_J\lambda_2 = \lambda_1^{\max\{1,2r\}}$ for some $r \in (0,1]$, then with confidence at least $1-\delta$, there holds*

$$\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D} \leq \left(\frac{2\mathcal{B}_{|D|,\lambda_1} \log\frac{2}{\delta}}{\sqrt{\lambda_1}} + 1 \right)^2.$$

Proof. We apply (4.5) to the operator $A = L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D$ and $B = L_K + \lambda_1 I + \lambda_2 J_D^T J_D$. Since

$$\|L_1 L_2\| = \|(L_1 L_2)^T\| = \|L_2^T L_1^T\| = \|L_2 L_1\| \tag{4.6}$$

for any self-adjoint operators L_1, L_2 on Hilbert spaces, applying $\|(\lambda_1 I + \lambda_2 J_D^T J_D + L_{K,D})^{-1}\| \leq \frac{1}{\lambda_1}$ and $\|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}\| \leq \frac{1}{\sqrt{\lambda_1}}$, we obtain

$$\begin{aligned} \mathcal{Q}_{D,\lambda_1,\lambda_2,J_D} &= \|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)(\lambda_1 I + \lambda_2 J_D^T J_D + L_{K,D})^{-1}\| \\ &\leq \|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\|^2 \lambda_1^{-1} \\ &\quad + \|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\| \lambda_1^{-\frac{1}{2}} + 1. \end{aligned}$$

To bound $\|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\|$, Proposition 4.1 implies almost surely

$$\begin{aligned} &\|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\| \\ &= \|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}(\lambda_1 + L_K)^{\frac{1}{2}}(\lambda_1 + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\| \end{aligned}$$

$$\leq \|(\lambda_1 I + \lambda_2 J_D^T J_D + L_K)^{-\frac{1}{2}}(\lambda_1 + L_K)^{\frac{1}{2}}\|\|(\lambda_1 + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\|$$
$$\leq 2\|(\lambda_1 + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\|.$$

Then by Lemma 4.2, we have

$$\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D} \leq \left(\frac{2\mathcal{B}_{|D|,\lambda_1}\log\frac{2}{\delta}}{\sqrt{\lambda_1}}\right)^2 + \left(\frac{2\mathcal{B}_{|D|,\lambda_1}\log\frac{2}{\delta}}{\sqrt{\lambda_1}}\right) + 1 \leq \left(\frac{2\mathcal{B}_{|D|,\lambda_1}\log\frac{2}{\delta}}{\sqrt{\lambda_1}} + 1\right)^2,$$

which finishes the proof of Proposition 4.3. □

4.2. Error decomposition

A preferable error decomposition in learning theory is to insert the data-free limitation of the studied algorithm, just as [6,17] did for RLS. However, for the MP-RLS algorithm, it is difficult to derive its data-free limitation, since the limitation of J_D is unclear, or even non-existent. Hence, we adopt a novel error decomposition for MP-RLS depended on Proposition 4.1. Let

$$f_{\lambda_1} := (L_K + \lambda_1 I)^{-1} L_K f_\rho$$

be the data free limitation of RLS. Then

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \leq \|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\|_\rho + \|f_{\lambda_1} - f_\rho\|_\rho, \tag{4.7}$$

where the first and second terms are called the sample and approximation errors, respectively. The following Proposition 4.4 provides the detailed error decomposition for MP-RLS.

Proposition 4.4. *Under Assumption 4, if* $\lambda_1, \lambda_2 \in (0,1)$ *satisfy* $2c_J\lambda_2 = \lambda_1^{\max\{1,2r\}}$ *for some* $r \in (0,1]$, *then there holds almost surely*

$$\begin{aligned}\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \leq\ & \mathcal{Q}_{D,\lambda_1,\lambda_2,J_D}\left(\|(L_K + \lambda_1 I)^{-\frac{1}{2}}\Delta'_D\|_K + \|(L_K + \lambda_1 I)^{-\frac{1}{2}}\Delta''_D\|_K\right)\\ &+\sqrt{2\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D}c_J\lambda_2}\|f_{\lambda_1}\|_K + \|f_{\lambda_1} - f_\rho\|_\rho\end{aligned} \tag{4.8}$$

where

$$\Delta'_D = S_D^T y_D - L_{K,D} f_\rho = \frac{1}{|D|}\sum_{(x,y)\in D}(y - f_\rho(x))K_x$$

and

$$\Delta''_D = L_{K,D}(f_\rho - f_{\lambda_1}) - L_K(f_\rho - f_{\lambda_1}) = \frac{1}{|D|}\sum_{(x,y)\in D}(f_\rho(x) - f_{\lambda_1}(x))K_x - L_K(f_\rho - f_{\lambda_1}).$$

Proof. Since $f_{\lambda_1} = (L_K + \lambda_1 I)^{-1} L_K f_\rho$, we have $\lambda_1 f_{\lambda_1} = L_K f_\rho - L_K f_{\lambda_1}$. Then

$$\begin{aligned} f_{D,\lambda_1,\lambda_2} - f_{\lambda_1} &= (L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1} S_D^T y_D - f_{\lambda_1}\\ &= (L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}(S_D^T y_D - L_{K,D} f_{\lambda_1} - L_K f_\rho + L_K f_{\lambda_1})\\ &\quad -(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}(\lambda_2 J_D^T J_D) f_{\lambda_1}\\ &= (L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\Delta'_D + (L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\Delta''_D\\ &\quad -(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}(\lambda_2 J_D^T J_D) f_{\lambda_1}. \end{aligned}$$

Therefore,

$$\|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\|_\rho \le I_1 + I_2 + I_3 \tag{4.9}$$

with

$$I_1 := \|(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\Delta_D'\|_\rho, \qquad I_2 := \|(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\Delta_D''\|_\rho$$

and

$$I_3 := \|(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}(\lambda_2 J_D^T J_D) f_{\lambda_1}\|_\rho.$$

The Cordes inequality [13] tells us that

$$\|A^s B^s\| \le \|AB\|^s \tag{4.10}$$

for positive operators A and B. Then Proposition 4.1 and (4.6) yield almost surely

$$\begin{aligned}
I_1 &= \|L_K^{\frac{1}{2}}(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\Delta_D'\|_K \\
&\le \|(L_K + \lambda_1 I)^{\frac{1}{2}}(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\Delta_D'\|_K \\
&= \|(L_K + \lambda_1 I)^{\frac{1}{2}}(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{\frac{1}{2}} \\
&(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{\frac{1}{2}} \\
&(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}(L_K + \lambda_1 I)^{\frac{1}{2}}(L_K + \lambda_1 I)^{-\frac{1}{2}}\Delta_D'\|_K \\
&\le 2\|(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\|\|(L_K + \lambda_1 I)^{-\frac{1}{2}}\Delta_D'\|_K \\
&= 2\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D}\|(L_K + \lambda_1 I)^{-\frac{1}{2}}\Delta_D'\|_K.
\end{aligned} \tag{4.11}$$

The same method as above derives follows almost surely

$$I_2 \le 2\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D}\|(L_K + \lambda_1 I)^{-\frac{1}{2}}\Delta_D''\|_K. \tag{4.12}$$

To bound I_3, since $2c_J\lambda_2 = \lambda_1^{\max\{1,2r\}}$ with $0 < \lambda_1 < 1$, using Proposition 4.1 and (4.10), we have almost surely

$$\begin{aligned}
I_3 &\le \|(L_K + \lambda_1 I)^{\frac{1}{2}}(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-1}\lambda_2 J_D^T J_D f_{\lambda_1}\|_K \\
&\le \|(L_K + \lambda_1 I)^{\frac{1}{2}}(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}\| \times \|(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{\frac{1}{2}} \\
&(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}\|\|(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}\|\|\lambda_2 J_D^T J_D\|\|f_{\lambda_1}\|_K \\
&\le \sqrt{2}\|(L_K + \lambda_1 I + \lambda_2 J_D^T J_D)^{\frac{1}{2}}(L_{K,D} + \lambda_1 I + \lambda_2 J_D^T J_D)^{-\frac{1}{2}}\|\lambda_1^{-1/2} c_J\lambda_2\|f_{\lambda_1}\|_K \\
&\le \sqrt{2c_J\lambda_2\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D}}\|f_{\lambda_1}\|_K.
\end{aligned} \tag{4.13}$$

Plugging (4.11), (4.12) and (4.13) into (4.9) and noting (4.7), we finish the proof of Proposition 4.4. □

5. Proof of main results

In this section, we aim at proving the main results.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

5.1. Proof of error bounds for MP-RLS

To prove Theorem 2.2, the following two lemmas which can be found in [6] and [17] are needed.

Lemma 5.1. *Let $0 < \delta < 1$ and D be a sample drawn independently according to ρ. Under Assumption 1 with $\frac{1}{2} \le r \le 1$, we have*

$$\|(L_K + \lambda_1 I)^{-\frac{1}{2}} \Delta'_D\|_K \le (\kappa M + \gamma)\mathcal{B}_{|D|,\lambda_1} \log\big(2/\delta\big). \tag{5.1}$$

Lemma 5.2. *Let $0 < \delta < 1$ and D be a sample drawn independently according to ρ and g be a measurable bounded function on $\mathcal{Z}$ and ξ_g be the random variable with values on $\mathcal{H}_K$ given by $\xi_g(z) = g(z)K_x$ for $z = (x, y) \in \mathcal{Z}$. With confidence at least $1 - \delta$, there holds*

$$\left\| (L_K + \lambda_1 I)^{-\frac{1}{2}} \left(\frac{1}{|D|} \sum_{z \in D} \xi_g(z) - E\left[\xi_g\right] \right) \right\|_K \le \|g\|_\infty \mathcal{B}_{|D|,\lambda_1} \log\big(2/\delta\big).$$

Based on Proposition 4.4, Proposition 4.3, Lemma 5.1 and Lemma 5.2, we can deduce the following proposition.

Proposition 5.3. *Let $0 < \delta < 1$. Under Assumptions 1–4 with $0 < r \le 1$ and $0 < \beta \le 1$, if $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_J\lambda_2 = \lambda_1^{\max\{1,2r\}}$, then with confidence at least $1 - \delta$, there holds*

$$\|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\|_\rho \le \left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1 \right)^2 \left(\log \frac{6}{\delta} \right)^3$$
$$\times \left[(\max\{\kappa M + \gamma, 2M\} + \|f_\rho - f_{\lambda_1}\|_\infty)\mathcal{B}_{|D|,\lambda_1} + \sqrt{2c_J\lambda_2}\|f_{\lambda_1}\|_K \right].$$

Proof. By Proposition 4.3, there exists a subset U_1 of $\mathcal{Z}^{|D|}$ with measure at least $1 - \frac{\delta}{3}$ such that for all $D \in U_1$

$$\mathcal{Q}_{D,\lambda_1,\lambda_2,J_D} \le \left(\frac{2\mathcal{B}_{|D|,\lambda_1} \log \frac{6}{\delta}}{\sqrt{\lambda_1}} + 1 \right)^2.$$

If Assumption 1 holds with $r \ge \frac{1}{2}$, then we have from Lemma 5.1 that there exists a subset U_2 of $\mathcal{Z}^{|D|}$ with measure at least $1 - \frac{\delta}{3}$ such that for all $D \in U_2$

$$\|(L_K + \lambda_1 I)^{-\frac{1}{2}} \Delta'_D\|_K \le (\kappa M + \gamma)\mathcal{B}_{|D|,\lambda_1} \log\big(6/\delta\big).$$

Otherwise, applying Lemma 5.2 to $\xi_g(z) = (y - f_\rho(x))K_x$, there exists a subset U_2 of $\mathcal{Z}^{|D|}$ with measure at least $1 - \frac{\delta}{3}$ such that for all $D \in U_2$

$$\left\| (L_K + \lambda_1 I)^{-\frac{1}{2}} \Delta'_D \right\|_K \le 2M\mathcal{B}_{|D|,\lambda_1} \log\big(6/\delta\big).$$

Finally, applying Lemma 5.2 to $\xi_g(z) = (f_\rho(x) - f_{\lambda_1}(x))K_x$, there exists a subset U_3 of $\mathcal{Z}^{|D|}$ with measure at least $1 - \frac{\delta}{3}$ such that for all $D \in U_3$

$$\left\| (L_K + \lambda_1 I)^{-\frac{1}{2}} \Delta''_D \right\|_K \le \|f_\rho - f_{\lambda_1}\|_\infty \mathcal{B}_{|D|,\lambda_1} \log\big(6/\delta\big).$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Combining the above estimates for the three terms of (4.8), the desired error bound holds true. □

The term $\|f_{\lambda_1} - f_\rho\|_\rho$ is independent of sample, and can be easily estimated as follows [29].

Lemma 5.4. *Under Assumption 1 with $0 < r \leq 1$, we have with confidence at least $1-\delta$*

$$\|f_{\lambda_1} - f_\rho\|_\rho \leq \lambda_1^r \|h_\rho\|_\rho \tag{5.2}$$

We see from Proposition 5.3 that the upper bound of $\|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\|_\rho$ depends on $\|f_{\lambda_1}\|_K$ and $\|f_\rho - f_{\lambda_1}\|_\infty$. The following Lemma 5.5 which can be found in [29] presents the bounds.

Lemma 5.5. *Under Assumption 1 with $0 < r \leq 1$ and Assumption 2, we have*

$$\|f_{\lambda_1}\|_K = \begin{cases} \lambda_1^{r-\frac{1}{2}} \|h_\rho\|_\rho, & \text{if } 0 < r < \frac{1}{2}, \\ \kappa^{2r-1} \|h_\rho\|_\rho & \text{if } \frac{1}{2} \leq r \leq 1 \end{cases} \tag{5.3}$$

and

$$\|f_\rho - f_{\lambda_1}\|_\infty = \begin{cases} M + \kappa \lambda_1^{r-\frac{1}{2}} \|h_\rho\|_\rho, & \text{if } 0 < r < \frac{1}{2}, \\ \kappa \lambda_1^{r-\frac{1}{2}} \|h_\rho\|_\rho & \text{if } \frac{1}{2} \leq r \leq 1. \end{cases} \tag{5.4}$$

By the help of the above preliminaries, we are in a position to prove Theorem 2.2.

Proof of Theorem 2.2. We divide the proof into two cases according to different regularity levels.

Case 1: $r \in (0, \frac{1}{2})$**.** It follows from Lemma 5.5 and Proposition 5.3 that

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \leq \left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1 \right)^2 \left(\log \frac{6}{\delta} \right)^3$$
$$\times \left[(3M + \kappa \lambda_1^{r-\frac{1}{2}}) \mathcal{B}_{|D|,\lambda_1} + \sqrt{2c_J \lambda_2} \lambda_1^{r-\frac{1}{2}} \|h_\rho\|_\rho \right] + \lambda^r \|h_\rho\|_\rho$$

holds with confidence at least $1-\delta$ for any $\delta \in (0,1)$. In this case, we choose $\lambda_1 = |D|^{-\frac{1}{1+\beta}}$, and $\lambda_2 = \frac{1}{2c_J}\lambda_1 = \frac{1}{2c_J}|D|^{-\frac{1}{1+\beta}}$ to obtain

$$\mathcal{B}_{|D|,\lambda_1} = \frac{2\kappa}{\sqrt{|D|}} \left\{ \frac{\kappa}{\sqrt{|D|\lambda_1}} + \sqrt{\mathcal{N}(\lambda_1)} \right\} \leq 2\kappa(\kappa + \sqrt{C_0}) |D|^{-\frac{1}{2(1+\beta)}}$$

and

$$\frac{\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} \leq 2\kappa(\kappa + \sqrt{C_0}).$$

Hence, the triangle inequality together with Lemma 5.4 shows that with confidence at least $1-\delta$, there holds

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \leq C_1 |D|^{-\frac{r}{1+\beta}} \left(\log \frac{6}{\delta} \right)^3,$$

where

$$C_1 := \left[4\kappa(\kappa + \sqrt{C_0}) + 1 \right]^2 \left[2\kappa(\kappa + \sqrt{C_0})(3M + \kappa) + \|h_\rho\|_\rho \right] + \|h_\rho\|_\rho.$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Case 2: $r \in [\frac{1}{2}, 1]$**.** It follows from Lemma 5.5 and Proposition 5.3 that for any $\delta \in (0,1)$, with confidence at least $1-\delta$, there holds

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \le \left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \left(\log\frac{6}{\delta}\right)^3 \times \left[(\kappa M + \gamma + \kappa\lambda_1^{r-1/2}\|h_\rho\|_\rho)\mathcal{B}_{|D|,\lambda_1} + \sqrt{2c_J\lambda_2}\kappa^{2r-1}\|h_\rho\|_\rho\right] + \lambda_1^r\|h_\rho\|_\rho.$$

Setting $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$ and $\lambda_2 = \frac{1}{2c_J}\lambda_1^{2r} = \frac{1}{2c_J}|D|^{-\frac{2r}{2r+\beta}}$, we have from Assumption 3 that

$$\mathcal{B}_{|D|,\lambda_1} = \frac{2\kappa}{\sqrt{|D|}}\left\{\frac{\kappa}{\sqrt{|D|\lambda_1}} + \sqrt{\mathcal{N}(\lambda_1)}\right\} \le 2\kappa(\kappa + \sqrt{C_0})|D|^{-\frac{r}{2r+\beta}}$$

and

$$\frac{\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} \le 2\kappa(\kappa + \sqrt{C_0}).$$

This together with Lemma 5.4 and the triangle inequality implies that

$$\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho \le C_2|D|^{-\frac{r}{2r+\beta}}\left(\log\frac{6}{\delta}\right)^3$$

holds with confidence at least $1-\delta$, where

$$C_2 := \|h_\rho\|_\rho + \left[4\kappa(\kappa + \sqrt{C_0}) + 1\right]^2 \times \left[2\kappa(\kappa M + \gamma + \kappa\|h_\rho\|_\rho)(\kappa + \sqrt{C_0}) + \kappa^{2r-1}\|h_\rho\|_\rho\right].$$

We write the above two upper bounds in a uniform way and take $\tilde{C} = \max\{C_1, C_2\}$ to complete the proof of (2.5). Applying the formula

$$E[\xi] = \int_0^\infty \mathrm{Prob}[\xi > t]dt \tag{5.5}$$

for nonnegative random variables to $\xi = \|f_{D,\lambda_1,\lambda_2} - f_\rho\|^2$, we get from (2.5) that

$$E[\|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho^2] \le \tilde{C}^2(\log^6 6 + 6\Gamma(7))|D|^{-\frac{2r}{\max\{2r,1\}+\beta}},$$

which finishes the proof of (2.6). The proof of Theorem 2.2 is completed. □

5.2. Proof of error bounds for distributed MP-RLS

In this part, we devote to proving Theorem 2.1. The main tool is the following error decomposition for the distributed algorithm (1.2), which can be found in [15].

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Proposition 5.6. *Let $\overline{f}_{D,\lambda_1,\lambda_2}$ be defined by (1.2), we have*

$$E\left[\|\overline{f}_{D,\lambda_1,\lambda_2}-f_\rho\|_\rho^2\right] \le 2\sum_{j=1}^m \frac{|D_j|^2}{|D|^2} E\left[\|f_{D_j,\lambda_1,\lambda_2}-f_{\lambda_1}\|_\rho^2\right] + 2\sum_{j=1}^m \frac{|D_j|}{|D|}\|E[f_{D_j,\lambda_1,\lambda_2}]-f_{\lambda_1}\|_\rho^2 + 2\|f_{\lambda_1}-f_\rho\|_\rho^2.$$

Before proving Theorem 2.1, we present a general error estimate for algorithm (1.2) without Assumption 3.

Theorem 5.7. *Under Assumption 1 with $\frac{1}{2}\le r\le 1$, Assumption 2 and Assumption 4, if $\lambda_1,\lambda_2\in(0,1)$ satisfy $2c_J\lambda_2=\lambda_1^{2r}$, then there exists a constant C independent of m or $|D_j|$ such that,*

$$E\left[\|\overline{f}_{D,\lambda_1,\lambda_2}-f_\rho\|_\rho^2\right] \le C\sum_{j=1}^m \frac{|D_j|^2}{|D|^2}\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^4\left[\frac{|D_j|}{|D|}\mathcal{B}_{|D_j|,\lambda_1}^2+\frac{\lambda_1^{2r-1}}{|D_j|}+\lambda_1^{2r}\right]. \tag{5.6}$$

Proof. As Assumption 1 holds with $\frac{1}{2}\le r\le 1$, it follows from Lemma 5.5 that $\|f_{\lambda_1}\|_K\le\kappa^{2r-1}\|h_\rho\|_\rho$ and $\|f_\rho-f_{\lambda_1}\|_\infty\le\kappa\lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho$. Then applying Proposition 5.3 to each data set D_j with $j=1,\cdots,m$, with confidence at least $1-\delta$, there holds

$$\|f_{D_j,\lambda_1,\lambda_2}-f_{\lambda_1}\|_\rho \le \left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2\left(\log\frac{6}{\delta}\right)^3 \times\left[(\kappa M+\gamma+\kappa\lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho)\mathcal{B}_{|D_j|,\lambda_1}+\kappa^{2r-1}\|h_\rho\|_\rho\sqrt{2c_J\lambda_2}\right].$$

Applying the formula (5.5) for nonnegative random variables to $\xi=\|f_{D_j,\lambda_1,\lambda_2}-f_{\lambda_1}\|^2$, we have

$$E\left[\|f_{D_j,\lambda_1,\lambda_2}-f_{\lambda_1}\|_\rho^2\right] \le (6\Gamma(7)+\log^6 6)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^4 \times\left[(\kappa M+\gamma+\kappa\lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho)\mathcal{B}_{|D_j|,\lambda_1}+\kappa^{2r-1}\|h_\rho\|_\rho\sqrt{2c_J\lambda_2}\right]^2. \tag{5.7}$$

Now we consider the term $\|E[f_{D_j,\lambda_1,\lambda_2}]-f_{\lambda_1}\|_\rho$. By Jensen's inequality, we have

$$\|E[f_{D_j,\lambda_1,\lambda_2}]-f_{\lambda_1}\|_\rho \le E\left[\|E^*[f_{D_j,\lambda_1,\lambda_2}]-f_{\lambda_1}\|_\rho\right],$$

where $E^*[f_{D_j,\lambda_1,\lambda_2}]$ denotes the conditional expectation w.r.t. y given $D_j(x)$. Therefore,

$$E^*[f_{D_j,\lambda_1,\lambda_2}]=(\lambda_1 I+L_{K,D_j}+\lambda_2 J_{D_j}^T J_{D_j})^{-1}L_{K,D_j}f_\rho,$$

which together with $\lambda_1 f_{\lambda_1}=L_K f_\rho-L_K f_{\lambda_1}$ yields

$$\begin{aligned}
&E^*[f_{D_j,\lambda_1,\lambda_2}]-f_{\lambda_1}=(\lambda_1 I+L_{K,D_j}+\lambda_2 J_{D_j}^T J_{D_j})^{-1}L_{K,D_j}f_\rho-f_{\lambda_1}\\
&=(\lambda_1 I+L_{K,D_j}+\lambda_2 J_{D_j}^T J_{D_j})^{-1}(L_{K,D_j}(f_\rho-f_{\lambda_1})-L_K(f_\rho-f_{\lambda_1}))\\
&\quad-(\lambda_1 I+L_{K,D_j}+\lambda_2 J_{D_j}^T J_{D_j})^{-1}\lambda_2 J_{D_j}^T J_{D_j}f_{\lambda_1}\\
&=(\lambda_1 I+L_{K,D_j}+\lambda_2 J_{D_j}^T J_{D_j})^{-1}\Delta''_{D_j}-(\lambda_1 I+L_{K,D_j}+\lambda_2 J_{D_j}^T J_{D_j})^{-1}\lambda_2 J_{D_j}^T J_{D_j}f_{\lambda_1}.
\end{aligned}$$

It follows from (4.12) and (4.13) with $D = D_j$ that there holds almost surely

$$\|(\lambda_1 I + L_{K,D_j} + \lambda_2 J_{D_j}^T J_{D_j})^{-1}\Delta''_{D_j}\|_\rho \le 2\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}}\|(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta''_{D_j}\|_K$$

and

$$\|(\lambda_1 I + L_{K,D_j} + \lambda_2 J_{D_j}^T J_{D_j})^{-1}\lambda_2 J_{D_j}^T J_{D_j} f_{\lambda_1}\|_\rho \le \sqrt{2\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}} c_J\lambda_2}\kappa^{2r-1}\|h_\rho\|_\rho.$$

Then by the Schwarz inequality we have

$$\begin{aligned} E\left[\|E^*[f_{D_j,\lambda_1,\lambda_2}] - f_{\lambda_1}\|_\rho\right] &\le 2E\left[\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}}\|(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta''_{D_j}\|_K\right] \\ &\quad + E\left[\sqrt{2\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}} c_J\lambda_2}\kappa^{2r-1}\|h_\rho\|_\rho\right] \\ &\le 2\left(E\left[\mathcal{Q}^2_{D_j,\lambda_1,\lambda_2,J_{D_j}}\right]\right)^{\frac{1}{2}}\left(E\left[\|(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta''_{D_j}\|_K^2\right]\right)^{\frac{1}{2}} \\ &\quad + \left(E\left[\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}}\right]\right)^{\frac{1}{2}}\sqrt{2c_J\lambda_2}\kappa^{2r-1}\|h_\rho\|_\rho. \end{aligned}$$

Applying the formula (5.5) to the nonnegative random variables $\xi_1 = \mathcal{Q}^2_{D_j,\lambda_1,\lambda_2,J_{D_j}}$ and $\xi_2 = \mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}}$ respectively, we have from Proposition 4.3 that

$$\begin{aligned} &E[\mathcal{Q}^2_{D_j,\lambda_1,\lambda_2,J_{D_j}}] = \int_0^\infty \mathrm{Prob}(\mathcal{Q}^2_{D_j,\lambda_1,\lambda_2,J_{D_j}} > t)dt = \int_0^\infty \mathrm{Prob}(\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}} > t^{\frac{1}{2}})dt \\ &\le (2\Gamma(5) + \log^4 2)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4 \end{aligned}$$

and

$$E[\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}}] = \int_0^\infty \mathrm{Prob}(\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}} > t)dt \le (4+\log^2 2)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2.$$

Therefore, we have from Lemma 5.4 and the proof of Proposition 18 in [17] that

$$\begin{aligned} &\|E[f_{D_j,\lambda_1,\lambda_2}] - f_{\lambda_1}\|_\rho^2 \le \left(E\left[\|E^*[f_{D_j,\lambda_1,\lambda_2}] - f_{\lambda_1}\|_\rho\right]\right)^2 \\ &\le 8\left(E\left[\mathcal{Q}^2_{D_j,\lambda_1,\lambda_2,J_{D_j}}\right]\right)\left(E\left[\|(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta''_{D_j}\|_K^2\right]\right) + 4c_J\lambda_2 E\left[\mathcal{Q}_{D_j,\lambda_1,\lambda_2,J_{D_j}}\right]\kappa^{4r-2}\|h_\rho\|_\rho^2 \\ &\le (16\Gamma(5) + 8\log^4 2)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4 \frac{\kappa^2\lambda_1^{2r-1}}{|D_j|} \\ &\quad + (4+\log^2 2)2c_J\lambda_2\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2\kappa^{4r-2}\|h_\rho\|_\rho^2. \end{aligned} \tag{5.8}$$

Inserting (5.7), (5.8) and (5.2) into Proposition 5.6 and noting $2c_J\lambda_2 = \lambda_1^{2r}$, we get

$$\begin{aligned} &E\left[\|\overline{f}_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] \le 2\sum_{j=1}^m \frac{|D_j|^2}{|D|^2}(6\Gamma(7) + \log^6 6)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4 \\ &\left[(\kappa M + \gamma + \kappa\lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho)\mathcal{B}_{|D_j|,\lambda_1} + \kappa^{2r-1}\|h_\rho\|_\rho\sqrt{2c_J\lambda_2}\right]^2 \end{aligned}$$

$$+2\sum_{j=1}^{m}\frac{|D_j|}{|D|}\left\{(16\Gamma(5)+8\log^4 2)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^4\frac{\kappa^2\lambda_1^{2r-1}}{|D_j|}\right.$$

$$\left.+(4+\log^2 2)^4 2c_J\lambda_2\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2\kappa^{4r-2}\|h_\rho\|_\rho^2\right\}+2\lambda_1^{2r}\|h_\rho\|_\rho^2$$

$$\le C\sum_{j=1}^{m}\frac{|D_j|}{|D|}\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^4\left[\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda_1}+\frac{\lambda_1^{2r-1}}{|D_j|}+\lambda_1^{2r}\right],$$

where

$$C:=(32\Gamma(7)+2\log^6 6)(\kappa M+\gamma+\kappa\|h_\rho\|_\rho+\kappa^{2r-1}\|h_\rho\|_\rho)^2+2\kappa^2(16\Gamma(5)+8\log^4 2)$$
$$+(4+\log^2 2)\kappa^{4r-2}\|h_\rho\|_\rho^2+2\|h_\rho\|_\rho^2.$$

The proof of Theorem 5.7 is completed. □

By the help of the above theorem, we can prove Theorem 2.1 as follows.

Proof of Theorem 2.1. For $\frac{1}{2}\le r\le 1$, we choose $\lambda_1=|D|^{-\frac{1}{2r+\beta}}$, and $\lambda_2=\frac{1}{2c_J}|D|^{-\frac{2r}{2r+\beta}}$. Then it follows from Assumption 3, $m\le|D|^{\frac{2r-1}{2r+\beta}}$ and $|D_1|=|D_2|=\cdots=|D_m|$ that

$$\frac{\mathcal{N}(\lambda_1)}{\lambda_1|D_j|}\le C_0 m|D|^{\frac{1-2r}{2r+\beta}}\le C_0.$$

We also have, for each $j=1,\ldots,m$,

$$\frac{\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}=\frac{2\kappa}{\sqrt{\lambda_1|D_j|}}\left\{\frac{\kappa}{\sqrt{|D_j|\lambda_1}}+\sqrt{\mathcal{N}(\lambda_1)}\right\}\le 2\kappa(\kappa+\sqrt{C_0}),$$

$$\frac{|D_j|}{|D|}\mathcal{B}^2_{|D_j|,\lambda_1}\le 8\kappa^2\left(\frac{m\kappa^2}{|D|^2\lambda_1}+\frac{\mathcal{N}(\lambda_1)}{|D|}\right)\le 8\kappa^2(\kappa^2+C_0)^2|D|^{-\frac{2r}{2r+\beta}}$$

and

$$\frac{\lambda_1^{2r-1}}{|D_j|}=\frac{|D|^{-\frac{2r-1}{2r+\beta}}m}{|D|}\le|D|^{-1}.$$

Then by Theorem 5.7,

$$E[\|\overline{f}_{D,\lambda_1,\lambda_2}-f_\rho\|_\rho^2]\le\hat{C}|D|^{-\frac{2r}{2r+\beta}},$$

where

$$\hat{C}:=C\left[4\kappa(\kappa+\sqrt{C_0})+1\right]^4\left[8\kappa^2(\kappa^2+C_0)^2+3\right].$$

This completes the proof of Theorem 2.1. □

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

5.3. Proof of error bounds for distributed manifold regularization

In this part, we aim at proving Theorem 3.1. For this purpose, we first prove the following theorem.

Theorem 5.8. *Under Assumption 1 with $0 < r \leq 1$ and Assumption 2, if $\lambda_1, \lambda_2 \in (0,1)$ satisfy $4\omega\kappa^2\lambda_2 = \lambda_1^{\max\{1,2r\}}$, then there exists a constant C^* independent of m, $|D_j|$ or $|D_j^*|$ such that*

$$\begin{aligned} &E\left[\|\overline{f}^*_{D^*,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] \\ &\leq C^* \sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^4 \Big[\frac{|D_j^*|}{|D^*|}\Big((M+\|f_\rho - f_{\lambda_1}\|_\infty)\mathcal{B}_{|D_j^*|,\lambda_1} \\ &\quad + \max\{M+\gamma, 2M\}\mathcal{B}_{|D_j|,\lambda_1}\Big)^2 + \lambda_1^{\max\{1,2r\}}\|f_{\lambda_1}\|_K^2 + \frac{\lambda_1^{2r-1}}{|D_j^*|} + \lambda_1^{2r}\|h_\rho\|_\rho^2\Big]. \end{aligned}$$

Proof. We first consider the term involving y, i.e., $(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta'_{D_j^*}$. Recalling

$$\Delta'_{D_j^*} = S^T_{D_j^*} y_{D_j^*} - L_{K,D_j^*} f_\rho = (S^T_{D_j^*} y_{D_j^*} - L_K f_\rho) + (L_K f_\rho - L_{K,D_j^*} f_\rho)$$

and

$$S^T_{D_j^*} y_{D_j^*} - L_K f_\rho = \frac{1}{|D_j^*|}\sum_{(x,y)\in D_j^*} yK_x - L_K f_\rho = \frac{1}{|D_j|}\sum_{(x,y)\in D_j} yK_x - L_K f_\rho = S^T_{D_j} y_{D_j} - L_K f_\rho,$$

it follows from Lemma 5.1 and Lemma 5.2 that with confidence at least $1-\frac{\delta}{4}$

$$\|(L_K+\lambda_1 I)^{-\frac{1}{2}}(S^T_{D_j^*} y_{D_j^*} - L_K f_\rho)\| \leq \max\{M+\gamma, 2M\}\mathcal{B}_{|D_j|,\lambda_1}\log\frac{8}{\delta}$$

and

$$\|(L_K+\lambda_1 I)^{-\frac{1}{2}}(L_K f_\rho - L_{K,D_j^*} f_\rho)\| \leq M\mathcal{B}_{|D_j^*|,\lambda_1}\log\frac{8}{\delta}$$

which means that there exists a subset U_4 of $\mathcal{Z}^{|D_j^*|}$ with measure at least $1-\frac{\delta}{2}$, such that for all $D_j^* \in U_4$

$$\left\|(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta'_{D_j^*}\right\|_K \leq (\max\{M+\gamma,2M\}\mathcal{B}_{|D_j|,\lambda_1} + M\mathcal{B}_{|D_j^*|,\lambda_1})\log\frac{8}{\delta}.$$

By Proposition 4.3 and (3.2), there exists a subset U_5 of $\mathcal{Z}^{|D_j^*|}$ with measure at least $1-\frac{\delta}{4}$ such that for $D_j^* \in U_5$,

$$\mathcal{Q}_{D_j^*,\lambda_1,\lambda_2,J_{D^*}} \leq \left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}\log\frac{8}{\delta}}{\sqrt{\lambda_1}}+1\right)^2.$$

From Lemma 5.2, there exists a subset U_6 of $\mathcal{Z}^{|D_j^*|}$ with measure at least $1-\frac{\delta}{4}$ such that for $D_j^* \in U_6$,

$$\left\|(L_K+\lambda_1 I)^{-\frac{1}{2}}\Delta''_{D_j^*}\right\|_K \leq \|f_\rho - f_{\lambda_1}\|_\infty \mathcal{B}_{|D_j^*|,\lambda_1}\log(8/\delta).$$

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Thus, from Proposition 4.4 with $c_J = 2\omega\kappa^2$, for any $\delta \in (0,1)$, with confidence at least $1-\delta$, there holds

$$\|\overline{f}^*_{D_j^*,\lambda_1,\lambda_2} - f_{\lambda_1}\|_\rho \le \left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \left(\log\frac{8}{\delta}\right)^3$$
$$\times \left[(M + \|f_\rho - f_{\lambda_1}\|_\infty)\mathcal{B}_{|D_j^*|,\lambda_1} + \max\{M+\gamma, 2M\}\mathcal{B}_{|D_j|,\lambda_1} + \sqrt{4\omega\kappa^2\lambda_2}\|f_{\lambda_1}\|_K\right].$$

We use the formula (5.5) again to get

$$E\left[\|\overline{f}^*_{D_j^*,\lambda_1,\lambda_2} - f_{\lambda_1}\|_\rho^2\right] \le (8\Gamma(7) + 2\log^6 8)\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4$$
$$\times \left[(M + \|f_\rho - f_{\lambda_1}\|_\infty)\mathcal{B}_{|D_j^*|,\lambda_1} + \max\{M+\gamma, 2M\}\mathcal{B}_{|D_j|,\lambda_1} + \sqrt{4\omega\kappa^2\lambda_2}\|f_{\lambda_1}\|_K\right]^2.$$

Since $E[\overline{f}^*_{D_j^*,\lambda_1,\lambda_2}]$ is independent of y, we just replace the D_j with D_j^* in (5.8) and obtain

$$\|E[\overline{f}^*_{D_j^*,\lambda_1,\lambda_2}] - f_{\lambda_1}\|_\rho^2 \le (16\Gamma(5) + 8\log^4 2)\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4 \frac{\kappa^2\lambda_1^{2r-1}}{|D_j^*|}$$
$$+ (4 + \log^2 2)4\omega\kappa^2\lambda_2\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \kappa^{4r-2}\|h_\rho\|_\rho^2.$$

Then it follows from Proposition 5.6 and Lemma 5.4 that

$$E\left[\|\overline{f}^*_{D^*,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] \le 2\sum_{j=1}^m \frac{|D_j^*|^2}{|D^*|^2}(8\Gamma(7) + 2\log^6 8)\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \left(\log\frac{8}{\delta}\right)^3$$
$$\times \left[(M + \|f_\rho - f_{\lambda_1}\|_\infty)\mathcal{B}_{|D_j^*|,\lambda_1} + \max\{M+\gamma, 2M\}\mathcal{B}_{|D_j|,\lambda_1} + \sqrt{4\omega\kappa^2\lambda_2}\|f_{\lambda_1}\|_K\right]$$
$$+ 2\sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\Big[(16\Gamma(5) + 8\log^4 2)\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4 \frac{\kappa^2\lambda_1^{2r-1}}{|D_j^*|}$$
$$+ (4 + \log^2 2)4\omega\kappa^2\lambda_2\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \kappa^{4r-2}\|h_\rho\|_\rho^2\Big] + 2\lambda_1^{2r}\|h_\rho\|_\rho^2$$
$$\le C^* \sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4 \Big[\frac{|D_j^*|}{|D^*|}\Big((M + \|f_\rho - f_{\lambda_1}\|_\infty)\mathcal{B}_{|D_j^*|,\lambda_1}$$
$$+ \max\{M+\gamma, 2M\}\mathcal{B}_{|D_j|,\lambda_1}\Big)^2 + \lambda_1^{\max\{1,2r\}}\|f_{\lambda_1}\|_K^2 + \frac{\lambda_1^{2r-1}}{|D_j^*|} + \lambda_1^{2r}\|h_\rho\|_\rho^2\Big],$$

where

$$C^* := 2\max\{(16\Gamma(5) + 8\log^4 2)\kappa^2, (16\Gamma(7) + 4\log^8 6) + 2(4 + \log^2 2)\kappa^{4r-2}\|h_\rho\|_\rho^2, \|h_\rho\|_\rho^2\}.$$

This completes the proof of Theorem 5.8. □

At last, we turn to prove Theorem 3.1.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Proof of Theorem 3.1. We divide the proof into two cases: $r \in (0, 1/2)$ and $r \in [1/2, 1]$.

Case 1: $r \in (0, 1/2)$**.** From Lemma 5.5, it follows $\|f_{\lambda_1}\|_K \le \lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho$ and $\|f_\rho - f_{\lambda_1}\|_\infty \le M + \kappa\lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho$. Inserting the above estimate into Theorem 5.8, we get

$$E\left[\|\overline{f}^*_{D^*,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] \le C_1^* \sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4$$
$$\times\Big[\frac{|D_j^*|}{|D^*|}\lambda_1^{r-1/2}\mathcal{B}_{|D_j^*|,\lambda_1} + \mathcal{B}_{|D_j|,\lambda_1}\Big)^2 + \frac{\lambda_1^{2r-1}}{|D_j^*|} + \lambda_1^{2r}\|h_\rho\|_\rho^2\Big], \tag{5.9}$$

where $C_1^* := C^*\Big(2(2M + \kappa\|h_\rho\|_\rho)^2 + 2(\kappa M + \gamma)^2 + 4M^2 + 6\|h_\rho\|_\rho^2 + \kappa^2\Big)$. Let $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$ and $\lambda_2 = \frac{1}{4\omega\kappa^2}|D|^{-\frac{1}{2r+\beta}}$, it follows from (2.3), $|D^*| \ge |D|^{\frac{1+\beta}{2r+\beta}}$, $m \le |D^*|\,|D|^{-\frac{1+\beta}{2r+\beta}}$ and $|D_1^*| = |D_2^*| = \cdots = |D_m^*|$ that $\frac{\lambda_1^{2r-1}}{|D_j^*|} = \frac{m|D|^{-\frac{2r-1}{2r+\beta}}}{|D^*|} \le |D|^{-1} \le |D|^{-\frac{2r}{2r+\beta}}$ and

$$\frac{\mathcal{N}(\lambda_1)}{\lambda_1|D_j^*|} \le \frac{C_0 m|D|^{\frac{1+\beta}{2r+\beta}}}{|D^*|} \le C_0.$$

Thus, for each $j = 1, \ldots, m$, there holds

$$\frac{\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda}} = \frac{2\kappa}{\sqrt{\lambda_1|D_j^*|}}\left\{\frac{\kappa}{\sqrt{|D_j^*|\lambda_1}} + \sqrt{\mathcal{N}(\lambda_1)}\right\} \le 2\kappa(\kappa + \sqrt{C_0})$$

and

$$\frac{|D_j^*|}{|D^*|}\lambda_1^{2r-1}\mathcal{B}^2_{|D_j^*|,\lambda_1} = \frac{1}{m}\left(\frac{\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda}}\right)^2\lambda_1^{2r} \le 4\kappa^2(\kappa + \sqrt{C_0})^2|D|^{-\frac{2r}{2r+\beta}}.$$

Moreover, $r + \beta \ge \frac{1}{2}$ and $m \le |D|^{\frac{2r+2\beta-1}{2r+\beta}}$ yield

$$\frac{|D_j^*|}{|D^*|}\mathcal{B}^2_{|D_j|,\lambda_1} \le 4\kappa^2(\kappa + C_0)^2\frac{1}{m}\max\left\{\frac{1}{|D_j|^2\lambda_1}, \frac{\lambda_1^{-\beta}}{|D_j|}\right\}$$
$$= 4\kappa^2(\kappa + C_0)^2\max\left\{m|D|^{-2+\frac{1}{2r+\beta}}, |D|^{-\frac{2r}{2r+\beta}}\right\} \le 4\kappa^2(\kappa^2 + C_0)^2|D|^{-\frac{2r}{2r+\beta}}.$$

Putting all the above estimates back into (5.9) yields

$$E\left[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2\right] \le C_3|D|^{-\frac{2r}{2r+\beta}},$$

where $C_3 := C_1^*\left[4\kappa(\kappa + \sqrt{C_0}) + 1\right]^4\left[4\kappa^2(\kappa^2 + C_0)^2 + 3\right]$.

Case 2: $\frac{1}{2} \le r \le 1$**.** By Lemma 5.5, we have $\|f_{\lambda_1}\|_K \le \kappa^{2r-1}\|h_\rho\|_\rho$ and $\|f_\rho - f_{\lambda_1}\|_\infty \le \kappa\lambda_1^{r-\frac{1}{2}}\|h_\rho\|_\rho$. Putting these bounds into Theorem 5.8, we have

$$E\left[\|\overline{f}^*_{D^*,\lambda_1,\lambda_2} - f_\rho\|_\rho^2\right] \le C_2^* \sum_{j=1}^m \frac{|D_j^*|}{|D^*|}\left(\frac{2\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^4\Big[\frac{|D_j^*|}{|D^*|}\Big(\mathcal{B}_{|D_j^*|,\lambda_1} + \mathcal{B}_{|D_j|,\lambda_1}\Big)^2$$
$$+ \frac{\lambda_1^{2r-1}}{|D_j^*|} + \lambda_1^{2r}\|h_\rho\|_\rho^2\Big] \tag{5.10}$$

where $C_2^* := C^*\left[2(M + \kappa\|h_\rho\|_\rho)^2 + 2(\kappa M + \gamma)^2 + 4M^2 + 2\kappa^{2r-1}\|h_\rho\|_\rho^2 + \kappa^2 + \|h_\rho\|_\rho^2\right]$.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

Let $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$ and $\lambda_2 = \frac{1}{4\omega\kappa^2}|D|^{-\frac{2r}{2r+\beta}}$. It then follows from (2.3), $m \le |D^*|\,|D|^{-\frac{1+\beta}{2r+\beta}}$ and $|D_1^*| = |D_2^*| = \cdots = |D_m^*|$ that $\frac{\lambda_1^{2r-1}}{|D_j^*|} = \frac{m|D|^{-\frac{2r-1}{2r+\beta}}}{|D^*|} \le |D|^{-1}$, and

$$\frac{\mathcal{N}(\lambda_1)}{\lambda_1|D_j^*|} \le \frac{C_0 m|D|^{\frac{1+\beta}{2r+\beta}}}{|D^*|} \le C_0.$$

Thus, for each $j = 1, \ldots, m$, there holds

$$\frac{\mathcal{B}_{|D_j^*|,\lambda_1}}{\sqrt{\lambda}} = \frac{2\kappa}{\sqrt{\lambda_1|D_j^*|}}\left\{\frac{\kappa}{\sqrt{|D_j^*|\lambda_1}} + \sqrt{\mathcal{N}(\lambda_1)}\right\} \le 2\kappa(\kappa + \sqrt{C_0}).$$

Furthermore, since $\mathcal{B}^2_{|D_j^*|,\lambda_1} \le \mathcal{B}^2_{|D_j|,\lambda_1}$, $m \le |D|^{\frac{2r+2\beta-1}{2r+\beta}}$ implies

$$\begin{aligned}\frac{|D_j^*|}{|D^*|}\mathcal{B}^2_{|D_j^*|,\lambda_1} &\le \frac{|D_j^*|}{|D^*|}\mathcal{B}^2_{|D_j|,\lambda_1} \le 4\kappa^2(\kappa + C_0)^2\frac{1}{m}\max\left\{\frac{1}{|D_j|^2\lambda_1}, \frac{\lambda_1^{-\beta}}{|D_j|}\right\}\\ &= 4\kappa^2(\kappa + C_0)^2\max\left\{m|D|^{-2+\frac{1}{2r+\beta}}, |D|^{-\frac{2r}{2r+\beta}}\right\} \le 4\kappa^2(\kappa^2 + C_0)^2|D|^{-\frac{2r}{2r+\beta}}.\end{aligned}$$

Putting all the above estimates back into (5.10), we have

$$E\left[\|\overline{f}_{D^*,\lambda} - f_\rho\|_\rho^2\right] \le C_4|D|^{-\frac{2r}{2r+\beta}},$$

where $C_4 = C_2^*\left[4\kappa(\kappa + \sqrt{C_0}) + 1\right]^4\left[4\kappa^2(\kappa^2 + C_0)^2 + 3\right]$, which proves Theorem 3.1. □

References

[1] Abhishake, S. Sivananthan, Multi-penalty regularization in learning theory, J. Complexity 35 (2016) 141–165.
[2] F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, J. Complexity 23 (2007) 52–72.
[3] G. Blanchard, N. Krämer, Optimal learning rates for kernel conjugate gradient regression, Adv. Neural Inf. Process. Syst. (2010) 226–234.
[4] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2339–2434.
[5] X. Chang, S.B. Lin, D.X. Zhou, Distributed semi-supervised learning with kernel ridge regression, J. Mach. Learn. Res. 18 (46) (2017) 1–22.
[6] A. Caponnetto, E. DeVito, Optimal rates for the regularized least squares algorithm, Found. Comput. Math. 7 (2007) 331–368.
[7] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Appl. Anal. 8 (2010) 161–183.
[8] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, Cambridge, 2007.
[9] D. Düvelmeyer, B. Hofmann, A multi-parameter regularization approach for estimating parameters in jump diffusion processes, J. Inverse Ill-Posed Probl. 14 (2006) 861–880.
[10] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.
[11] Y. Feng, S.G. Lv, H. Hang, J.A.K. Suykens, Kernelized elastic net regularization: generalization bounds, and sparse recovery, Neural Comput. 28 (2016) 525–562.
[12] M. Fornasier, V. Naumova, S.V. Pereverzev, Parameter choice strategies for multi-penalty regularization, SIAM J. Numer. Anal. 52 (2014) 1170–1194.
[13] J. Fujii, M. Fujii, T. Furuta, R. Nakamoto, Norm inequalities equivalent to Heinz inequality, Proc. Amer. Math. Soc. 118 (1993) 827–830.
[14] Z.C. Guo, D.X. Zhou, Concentration estimates for learning with unbounded sampling, Adv. Comput. Math. 38 (2013) 207–223.
[15] Z.C. Guo, S.B. Lin, D.X. Zhou, Learning theory for distributed spectral algorithm, Inverse Probl. 33 (2017) 074009.
[16] T. Hu, J. Fan, Q. Wu, D.X. Zhou, Regularization schemes for minimum error entropy principle, Appl. Anal. 13 (2015) 437–455.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001

[17] S.B. Lin, X. Guo, D.X. Zhou, Distributed learning with regularized least squares, J. Mach. Learn. Res. (2016), Minor revision under review, arXiv:1608.03339v2.
[18] S.B. Lin, D.X. Zhou, Distributed kernel gradient descent algorithms, Constr. Approx. (2017), http://dx.doi.org/10.1007/s00365-017-9379-1.
[19] S. Lu, S.V. Pereverzev, U. Tautenhahn, Dual regularized total least squares and multi-parameter regularization, Comput. Methods Appl. Math. 8 (2008) 253–262.
[20] S. Lu, S.V. Pereverzev, Multi-parameter regularization and its numerical realization, Numer. Math. 118 (2011) 1–31.
[21] S. Lu, S. Pereverzyev Jr., S. Sivananthan, Multiparameter regularization for construction of extrapolating estimators in statistical learning theory, in: Multiscale Signal Analysis and Modeling, Springer, New York, 2013, pp. 347–366.
[22] Y. Lu, L. Shen, Y. Xu, Multi-parameter regularization methods for high-resolution image reconstruction with displacement errors, IEEE Trans. Circ. Syst. I Fund. Theory Appl. 54 (2007) 1788–1799.
[23] G. Mann, R. McDonald, M. Mohri, N. Silberman, D. Walker, Efficient large-scale distributed training of conditional maximum entropy models, Adv. Neural Inf. Process. Syst. (2009) 1231–1239.
[24] V. Naumova, S. Peter, Minimization of multi-penalty functionals by alternating iterative thresholding and optimal parameter choices, Inverse Probl. 30 (2014) 125003.
[25] V. Roth, The generalized LASSO, IEEE Trans. Neural Netw. 15 (2004) 16–28.
[26] L. Shi, Y.L. Feng, D.X. Zhou, Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces, Appl. Comput. Harmon. Anal. 31 (2011) 286–302.
[27] S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. 41 (2004) 279–305.
[28] S. Smale, D.X. Zhou, Shannon sampling II: connections to learning theory, Appl. Comput. Harmon. Anal. 19 (2005) 285–302.
[29] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.
[30] I. Steinwart, D. Hush, C. Scovel, Optimal rates for regularized least squares regression, COLT (2009).
[31] W. Wang, S. Lu, H. Mao, J. Cheng, Multi-parameter Tikhonov regularization with the ℓ^0 sparsity constraint, Inverse Probl. 29 (2013) 065018.
[32] Q. Wu, D.X. Zhou, Learning with sample dependent hypothesis space, Comput. Math. Appl. 56 (2008) 2896–2907.
[33] P. Xu, Y. Fukuda, Y. Liu, Multiple parameter regularization: numerical solutions and applications to the determination of geopotential from precise satellite orbits, J. Geod. 80 (2006) 17–27.
[34] Y.C. Zhang, J. Duchi, M. Wainwright, Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates, J. Mach. Learn. Res. 16 (2015) 3299–3340.
[35] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2005) 301–320.

Please cite this article in press as: Z.-C. Guo et al., Distributed learning with multi-penalty regularization, Appl. Comput. Harmon. Anal. (2017), http://dx.doi.org/10.1016/j.acha.2017.06.001