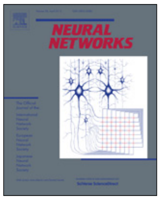
Neural Networks 44 (2013) 44–50

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Convergence rate of the semi-supervised greedy algorithm

Hong Chen [a,b,*], Yicong Zhou [b], Yuan Yan Tang [b], Luoqing Li [c], Zhibin Pan [a]

[a] *College of Science, Huazhong Agricultural University, Wuhan 430070, China*

[b] *Department of Computer and Information Science, University of Macau, Macau 999078, China*

[c] *Faculty of Mathematics and Computer Science, Hubei University, Wuhan 430062, China*

ARTICLE INFO

Article history:
Received 17 December 2011
Received in revised form 21 December 2012
Accepted 1 March 2013

Keywords:
Semi-supervised learning
Sparse
Greedy algorithm
Data-dependent hypothesis space
Generalization error

ABSTRACT

This paper proposes a new greedy algorithm combining the semi-supervised learning and the sparse representation with the data-dependent hypothesis spaces. The proposed greedy algorithm is able to use a small portion of the labeled and unlabeled data to represent the target function, and to efficiently reduce the computational burden of the semi-supervised learning. We establish the estimation of the generalization error based on the empirical covering numbers. A detailed analysis shows that the error has $O(n^{-1})$ decay. Our theoretical result illustrates that the unlabeled data is useful to improve the learning performance under mild conditions.

1. Introduction

The semi-supervised learning, i.e., learning from a set of the labeled and unlabeled data, has attracted many researchers recently due to its main challenge in how to improve its prediction performance using a few labeled data with a large set of unlabeled data. In literature, algorithms of the semi-supervised learning have been proposed in different perspectives. Examples include the graph-based learning (Belkin & Niyogi, 2004; Belkin, Niyogi, & Sindhwani, 2006; Chen, Li, & Peng, 2009; Johnson & Zhang, 2007, 2008), co-training (Blum & Mitchell, 1998; Sindhwani, Niyogi, & Belkin, 2005; Sindhwani & Rosenberg, 2008) and many others. A review study of the semi-supervised learning is discussed in Chapelle, Schölkopf, and Zien (2006) and Zhu (2005).

Among these methods proposed for semi-supervised learning, a family of them can be unified in a Tikhonov regularization scheme in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ with a Mercer kernel K, e.g., Belkin and Niyogi (2004), Sindhwani et al. (2005) and Sindhwani and Rosenberg (2008). For the labeled data $\{(x_i, y_i)\}_{i=1}^n$ and the unlabeled data $\{x_i\}_{i=n+1}^{n+m}$, the solution of the regularization framework usually has the following expression

$$\sum_{i=1}^{m+n} \alpha_i K(x_i, \cdot), \quad \alpha_i \in \mathbb{R}.$$

The semi-supervised algorithms intend to search the coefficients $\{\alpha_i\}$ for promising prediction performance. Although they are excellent in the empirical evaluation (Belkin & Niyogi, 2004; Sindhwani et al., 2005; Sindhwani & Rosenberg, 2008), two issues remain to be further addressed in theory:

- Computation difficulty. Because the regularized framework generally uses the kernel expansions of all the labeled and unlabeled data, computation becomes a serious problem for a huge set of the unlabeled data in real applications.
- Manifold assumption. In many graph-based methods such as Belkin and Niyogi (2004), Sindhwani et al. (2005) and Sindhwani and Rosenberg (2008), it is assumed that the high-dimensional data is relied on a low-dimensional manifold. However, for different types of data, the convincing evidences of the manifold structure are not available (Fan, Gu, Qiao, & Zhang, 2011).

To address the above issues, previous discussions have done to realize sparse semi-supervised learning in Fan et al. (2011), Sun and Shawe Taylor (2010) and Tsang and Kwok (2007) but the limitation is that they just use unlabeled data to construct an additional sparse regularization term.

In this paper, we investigate the sparse representation of the semi-supervised learning without manifold assumption, and consider the sparsity of the semi-supervised learning in data dependent hypothesis spaces. Inspired by the greedy algorithms in Barron, Cohen, Dahmen, and DeVore (2008), Nair, Choudhury, and Keane (2007) and Zhang (2002, 2009), we propose a new sparse greedy algorithm for the semi-supervised learning. Theoretical analysis shows that the proposed algorithm is efficient to realize

[*] Corresponding author at: College of Science, Huazhong Agricultural University, Wuhan 430070, China. Tel.: +86 18971089571.
E-mail addresses: chenhongml@163.com, chenh@mail.hzau.edu.cn (H. Chen).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.03.001

the sparse learning. Several contributions of this work have been highlighted below:

- Our method integrates three different machine learning methods in a coherent way: the sparse semi-supervised learning (Fan et al., 2011; Sun & Shawe Taylor, 2010; Tsang & Kwok, 2007), the greedy algorithm (Nair et al., 2007; Zhang, 2002, 2009), and the error analysis in data dependent hypothesis spaces (Shi, Feng, & Zhou, 2011; Sun & Wu, 2011; Wu & Zhou, 2008; Xiao & Zhou, 2010). We also show how to use them to design and analyze a new semi-supervised algorithm.
- Generalization error bounds are derived for nonsymmetric and indefinite kernels. Theoretical results show the relative values of the labeled data and unlabeled data to achieve fast learning rates. In particular, we illustrate that the role of the unlabeled data is twofold. The first one is that the semi-supervised method can achieve fast learning rates using the additionally unlabeled data. The second one is that the learning rates essentially depend on the number of the labeled data even if the number of unlabeled data tends to infinity. Furthermore, our error analysis results rely on weaker conditions than the previous methods which are based on density assumption or manifold assumption in Belkin et al. (2006), Belkin and Niyogi (2004), Chen and Li (2009), Chen et al. (2009), Chen, Li, and Peng (2010), Johnson and Zhang (2007, 2008) and Rigollet (2007).
- Even for the supervised learning settings, we can achieve faster learning rates than the previous results in Xiao and Zhou (2010), Shi et al. (2011) and Sun and Wu (2011). In particular, our analysis does not require the interior cone condition presented in Shi et al. (2011) and Xiao and Zhou (2010).

The organization of this paper is as follows. Section 2 provides the necessary background of the semi-supervised learning and then presents the sparse semi-supervised greedy algorithm. Section 3 includes the main result on error analysis and its proof is given in Section 4. An empirical study is given in Section 5. We conclude the paper in Section 6.

2. The sparse semi-supervised greedy algorithm

Let the input space $\mathcal{X} \subset \mathbb{R}^d$ be a compact subset and $\mathcal{Y} = [-M, M]$. In the semi-supervised model, a learner obtains a labeled data set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$ and an unlabeled data set $\mathbf{x} = \{x_{n+j}\}_{j=1}^m$. Here, the labeled examples $(x_i, y_i) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, $1 \leqslant i \leqslant n$, are independent copies of the random element (x, y) having distribution ρ on $\mathcal{Z}$. The unlabeled data x_{n+j}, $1 \leqslant j \leqslant m$, are independent copies of $\mathcal{X}$, whose distribution (the margin distribution of ρ) is denoted by $\rho_{\mathcal{X}}$. The learning goal is to pick up a function $f : \mathcal{X} \to \mathcal{Y}$ to minimize the expected error

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 d\rho.$$

The function that minimizes the error is called the regression function. It is given by

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

where $\rho(\cdot|x)$ is the conditional probability measure at x induced by ρ.

For the given training data $\mathbf{z}$, we define the empirical norm

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n |f(x_i)|^2.$$

Note that $\|\cdot\|_n$ is the $L^2_{\rho_{\mathcal{X}}}$ norm with respect to the discrete measure $\upsilon_{\mathbf{z}} := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$, where δ_u is the Dirac measure at u.

Denote the empirical error as

$$\mathcal{E}_{\mathbf{z}}(f) := \|f - y\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2.$$

We usually call a symmetric and positive semi-definite continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel. The RKHS $\mathcal{H}_K$ is defined to be the closure of the linear span of a set of functions $\{K_x := K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot \rangle_K$ given by $\langle K_x, K_{x'} \rangle_K = K(x, x')$. Using the form of the manifold regularization in the RKHS, a semi-supervised algorithm was proposed in Belkin and Niyogi (2004):

$$f_{\mathbf{z},\mathbf{x}} = \arg\min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda_1 \|f\|_K^2 + \lambda_2 \sum_{i,j=1}^{m+n} W_{ij} (f(x_i) - f(x_j))^2 \right\}, \tag{2.1}$$

where $\lambda_1, \lambda_2 > 0$, are the regularization parameters and W_{ij} is the similarity weight related to x_i and x_j.

The empirical evaluation in Belkin and Niyogi (2004) has shown the excellent performance of the semi-supervised algorithm in (2.1). However, the solution $f_{\mathbf{z},\mathbf{x}} = \sum_{i=1}^{m+n} \alpha_i K_{x_i}$ generally includes kernel expansions of all the labeled and unlabeled data. As mentioned in Tsang and Kwok (2007), this method may result in computation difficulty for a large set of the unlabeled data.

In this paper, we use a greedy algorithm to realize the sparse semi-supervised learning. Denote the hypothesis space (depending on $\mathbf{z}$ and $\mathbf{x}$) as

$$\mathcal{H}_{\mathbf{z},\mathbf{x}} = \left\{ \sum_{i=1}^{m+n} \alpha_i K_{x_i} : \alpha_i \in \mathbb{R} \right\}.$$

The ℓ_1 norm is defined as

$$\|f\|_{\ell_1} = \inf \left\{ \sum_{i=1}^{m+n} |\alpha_i| : f = \sum_{i=1}^{m+n} \alpha_i K_{x_i} \in \mathcal{H}_{\mathbf{z},\mathbf{x}} \right\}. \tag{2.2}$$

Different from the previous hypothesis spaces which are based on the Mercer kernel, we only require the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a continuous function. This means that K here is not necessarily symmetric or positive semi-definite. A wider selection of the kernel offers more flexibility. Explicit examples of this general kernel can be found in Shi et al. (2011) and Xiao and Zhou (2010).

The definition of f_ρ tells us $|f_\rho(x)| \leqslant M$, so it is natural to restrict the approximation functions to $[-M, M]$. The projection operator has been used in error analysis of learning algorithms (for example, Barron et al., 2008, Chen, Wu, Ying, & Zhou, 2004 and Cucker & Zhou, 2007).

Definition 1. The projection operator $\pi = \pi_M$ is defined on the space of the measurable functions $f : \mathcal{X} \to \mathcal{R}$ as

$$\pi(f)(x) = \begin{cases} M, & \text{if } f(x) > M; \\ -M, & \text{if } f(x) < -M; \\ f(x), & \text{otherwise.} \end{cases}$$

Since $(\pi(f)(x) - y)^2 \leqslant (f(x) - y)^2$, we know $\mathcal{E}(\pi(f)) \leqslant \mathcal{E}(f)$ and $\mathcal{E}_{\mathbf{z}}(\pi(f)) \leqslant \mathcal{E}_{\mathbf{z}}(f)$. Therefore, using $\pi(f)$ to estimate f_ρ is more accurate than using f. According to this, we introduce the projection operator in our algorithm.

We now present the sparse semi-supervised greedy algorithm in Table 2.1.

The algorithm can be divided into two parts. Selecting features ϕ_k and solving the empirical risk minimization to derive $\hat{f}_k$. The normalization of kernels makes error analysis feasible while maintaining the prediction performance of the algorithm. The stopping condition $\|y - \hat{f}_k\|_n^2 + \|\hat{f}_k\|_{\ell_1} \leqslant \|y\|_n^2$ is inspired from the initial function $\hat{f}_0 = 0$.

Table 2.1
Semi-supervised greedy algorithm.

```
Input: $\mathbf{z} \in \mathcal{Z}^n,\ \mathbf{x} \in \mathcal{X}^m$, $K$, and $T > 0$
  Step 1. Normalization: $\hat{K}_{x_i} = K_{x_i}/\|K_{x_i}\|_{m+n},\ i = 1, \ldots, m+n$
    Dictionary: $\mathcal{D}_{m+n} = \{\hat{K}_{x_i} : i = 1, \ldots, m+n\}$
  Step 2. Computation: Let $\hat{f}_0 = 0$
    for$k = 1, 2, \ldots$
      let $\phi_k = \arg\min_{g \in \mathcal{D}_{m+n}} |\langle y - \hat{f}_{k-1}, g\rangle_n|$
      let $\hat{\mathcal{H}}_k = \mathrm{Span}(\phi_1, \ldots, \phi_k)$
      $\hat{f}_k = \arg\min_{h \in \hat{\mathcal{H}}_k} \|y - h\|_n^2$
      if$\|y - \hat{f}_k\|_n^2 + \|\hat{f}_k\|_{\ell_1} \le \|y\|_n^2$ and $k \geqslant T$break
    end
Output: $\pi(\hat{f}_k)$
```

3. Main result

Now we introduce a data-free function space similar to Shi et al. (2011) and Xiao and Zhou (2010).

Definition 2. Define a data-free assumption function space

$$\mathcal{H}_1 = \left\{ f : f = \sum_{j=1}^{\infty} \alpha_j \tilde{K}_{u_j},\ \{\alpha_j\} \in \ell_1,\ \{u_j\} \subset \mathcal{X}, \right.$$
$$\left. \tilde{K}_{u_j} = K_{u_j}/\|K_{u_j}\|_{L^2_{\rho_\mathcal{X}}} \right\}$$

with the norm

$$\|f\|_{\mathcal{H}_1} = \inf\left\{ \sum_{j=1}^{\infty} |\alpha_j| : f = \sum_{j=1}^{\infty} \alpha_j \tilde{K}_{u_j} \right\}.$$

In order to investigate the approximation of $\pi(\hat{f}_k)$ to f_ρ, we introduce a regularizing function

$$f_\lambda = \arg\min_{f \in \mathcal{H}_1} \{\mathcal{E}(f) + \lambda\|f\|_{\mathcal{H}_1}\},$$

where $\lambda > 0$ is a regularization parameter.

The regularizing error can be expressed as

$$D(\lambda) = \inf_{f \in \mathcal{H}_1} \{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|_{\mathcal{H}_1}\}.$$

The decay of $D(\lambda)$ as $\lambda \to 0$ measures the approximation ability of the function space $\mathcal{H}_1$ to f_ρ. It is easy to see that

$$\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho) \leqslant \left\{ \mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}_{\mathbf{z}}(\pi(\hat{f}_k)) + \mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda) \right\}$$
$$+ \left\{ \mathcal{E}_{\mathbf{z}}(\pi(\hat{f}_k)) - \mathcal{E}_{\mathbf{z}}(f_\lambda) \right\} + \left\{ \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) \right\}.$$

In learning theory, three terms in the right part of the above inequality are called the sample, hypothesis, and approximation errors, respectively.

The following two conditions have been widely used for error analysis in extensive literature, e.g., Cucker and Zhou (2007), Shi et al. (2011), Wu and Zhou (2008), Xiao and Zhou (2010) and Zou, Li, and Xu (2009). They are also the necessary conditions for establishing our approximation analysis.

Definition 3. We say that the target function f_ρ can be approximated with exponent $0 < q \leqslant 1$ in $\mathcal{H}_1$ if there exists a constant $c_q \geqslant 1$, such that

$$D(\lambda) \leqslant c_q \lambda^q, \quad \forall \lambda > 0.$$

Definition 4. We say that the kernel function K is a C^s kernel with $s > 0$ if there exists some constant $c_s > 0$, such that

$$|K(t, x) - K(t, x')| \leqslant c_s |x - x'|^s, \quad \forall t, x, x' \in \mathcal{X}.$$

We now formulate the generalization error bounds for the semi-supervised algorithm defined in Table 2.1.

Theorem 1. *Assume that f_ρ can be approximated with exponent $0 < q \leqslant 1$ in $\mathcal{H}_1$ and K is a C^s kernel with $0 < k_0 \leqslant K(u, v) \leqslant k_1$ for any $u, v \in \mathcal{X}$. Choose $T \geqslant n$. Then, for any $0 < \delta < 1$, with confidence $1 - \delta$*

$$\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho) \leqslant \tilde{c} \log(1/\delta) n^{-\min\left\{\frac{2}{2+p}, \frac{q}{2-q}\right\}},$$

where constant $\tilde{c}$ is independent of m, k, δ, and

$$p = \begin{cases} 2d/(d+2s), & \text{if } 0 < s \leqslant 1; \\ 2d/(d+2), & \text{if } 1 < s \leqslant 1 + d/2; \\ d/s, & \text{if } s > 1 + d/2. \end{cases}$$

When $q \to 1, s \to \infty$, we can derive the learning rate of $O(n^{-1})$. The convergence rate is faster than the sparse semi-supervised method in Sun and Shawe Taylor (2010) with the order of $O(n^{-\frac{1}{2}})$. Although the convergence rates of the exponential order are presented for the semi-supervised algorithms in Chen and Li (2009) and Rigollet (2007), these results depend on the mixture density assumption or the strong cluster assumption.

Meanwhile, our convergence rate is faster than the supervised coefficient regularization methods, e.g., $O(n^{-\frac{1}{5}})$ in Sun and Wu (2011), $O(n^{-\frac{1}{3}})$ in Xiao and Zhou (2010), $O(n^{-\frac{1}{2}})$ in Shi et al. (2011). Although an additional lower bound of the kernel K is required in current analysis, we do not need the restricted conditions of $\mathcal{X}$ and ρ presented in Shi et al. (2011) and Xiao and Zhou (2010).

Our result also illustrates that the role of unlabeled data is twofold in theory: (1) the unlabeled data is useful to improve the learning performance of the greedy algorithm; and (2) its affect is limited by the present analysis framework because the sample error essentially depends on the number of the labeled data.

4. Error analysis

In this section, we provide the proof of Theorem 1 based on the upper bound analysis of the sample and hypothesis errors. The sample error is bounded by the error analysis method and empirical covering numbers. The hypothesis error is established in terms of theoretical analysis of the greedy algorithm presented in Barron et al. (2008).

4.1. Estimate of sample error

We first establish the estimation of the sample error using the standard analysis techniques presented in Cucker and Smale (2002), Cucker and Zhou (2007) and Shi et al. (2011). For completeness, we include them here with some proofs.

Denote

$$S_1 = \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)\}$$

and

$$S_2 = \{\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi(\hat{f}_k)) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}.$$

We can observe that the sample error

$$\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}_{\mathbf{z}}(\pi(\hat{f}_k)) + \mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda) = S_1 + S_2.$$

The bound of S_1 has been proved in Shi et al. (2011) using the one-side Bernstein inequality and the fact $\|f_\lambda\|_{\mathcal{H}_1} \leqslant D(\lambda)/\lambda$.

Proposition 1. *For any $\delta > 0$, with confidence at least $1 - \delta$, we have*

$$S_1 \leqslant \frac{7(3M + k_1 k_0^{-1} D(\lambda)/\lambda)}{3n} \log(1/\delta) + \frac{1}{2} D(\lambda).$$

Lemma 1. *Under the conditions of Theorem 1, for almost every $\mathbf{z} \in \mathcal{Z}^m$ and $\mathbf{x} \in \mathcal{X}^n$, we have*

$$\|\hat{f}_k\|_{\mathcal{H}_1} \leqslant k_1 M^2.$$

Proof. From the definition of the ℓ_1 norm in (2.2) and Definition 2, we know that

$$\|\hat{f}_k\|_{\mathcal{H}_1} \leqslant k_1 \|\hat{f}_k\|_{\ell_1}.$$

Based on the algorithm in Table 2.1, we observe that

$$\|\hat{f}_k\|_{\ell_1} \leqslant \|y\|_n^2 \leqslant M^2.$$

Combining above two inequalities, we derive the desired result. □

In order to obtain the uniform upper bound of S_2, we consider the data-dependent space

$$\mathcal{B}_r = \{f \in \mathcal{H}_1 : \|f\|_{\mathcal{H}_1} \leqslant r\},$$

where $r = k_1 M^2$.

Recently, a nice result has been established in Shi et al. (2011) to estimate the capacity of $\mathcal{B}_1$. Now recall some basic notations and definitions.

Definition 5. Let $(\mathcal{U}, d)$ be a pseudo-metric space and $S \subset \mathcal{U}$ denote a subset. For every $\epsilon > 0$, the covering number $\mathcal{N}(S, \epsilon, d)$ of S with respect to ϵ, d is defined as the minimal number of balls of radius ϵ whose union covers S, that is,

$$\mathcal{N}(S, \epsilon, d) = \min\left\{ l \in \mathbb{N} : S \subset \bigcup_{j=1}^{l} B(s_j, \epsilon) \right. \\ \left. \text{for some } \{s_j\}_{j=1}^{l} \subset \mathcal{U} \right\},$$

where $B(s_j, \epsilon) = \{s \in \mathcal{U} : d(s, s_j) \leqslant \epsilon\}$ is a ball in $\mathcal{U}$.

The empirical covering number with the ℓ_2 metric is defined below.

Definition 6. Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$, $\mathbf{u} = (x_i)_{i=1}^k$ and $\mathcal{F}|_{\mathbf{u}} = \{(f(u_i))_{i=1}^k : f \in \mathcal{F}\} \subset \mathbb{R}^k$. Set $\mathcal{N}_{2,\mathbf{u}}(\mathcal{F}, \epsilon) = \mathcal{N}_{2,\mathbf{u}}(\mathcal{F}|_{\mathbf{u}}, \epsilon, d_2)$. The ℓ_2 empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{k \in \mathbb{N}} \sup_{\mathbf{u} \in \mathcal{X}^k} \mathcal{N}_{2,\mathbf{u}}(\mathcal{F}, \epsilon), \quad \epsilon > 0,$$

where the ℓ_2 metric

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{k} \sum_{i=1}^{k} |a_i - b_i|^2 \right)^{\frac{1}{2}}, \quad \forall \mathbf{a} = (a_i)_{i=1}^k \in \mathbb{R}^k, \\ \mathbf{b} = (b_i)_{i=1}^k \in \mathbb{R}^k.$$

Now we introduce an important concentration inequality, which can be found in Wu, Ying, and Zhou (2007).

Lemma 2. *Assume that there are constants $B, c > 0$ and $\alpha \in [0, 1]$ such that $\|f\|_\infty \leqslant B$ and $Ef \leqslant c(Ef)^\alpha$ for every $f \in \mathcal{F}$. For some $a > 0$ and $p \in (0, 2)$, if*

$$\log(\mathcal{N}_2(\mathcal{F}, \epsilon)) \leqslant a\epsilon^{-p}, \quad \forall \epsilon > 0,$$

then there exists a constant c'_p depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$Ef - \frac{1}{m} \sum_{i=1}^{n} f(z_i) \leqslant \frac{1}{2} \eta^{1-\alpha} (Ef)^\alpha + c'_p \eta + 2 \left(\frac{ct}{n} \right)^{\frac{1}{2-\alpha}} \\ + \frac{18Bt}{n}, \quad \forall f \in \mathcal{F},$$

where

$$\eta := \max\left\{ c^{\frac{2-p}{4-2\alpha+p\alpha}} \left(\frac{a}{n} \right)^{\frac{2}{4-2\alpha+p\alpha}}, B^{\frac{2-p}{2+p}} \left(\frac{a}{n} \right)^{\frac{2}{2+p}} \right\}.$$

Proposition 2. *If K is a C^s kernel, then for any $0 < \delta < 1$, with confidence at least $1 - \delta$,*

$$S_2 \leqslant \frac{1}{2} \{\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho)\} + c_1 \log(1/\delta) n^{-\frac{2}{2+p}},$$

where $c_1 = 640M^2 (c_{p,K}(4Mr)^p)^{\frac{2}{2+p}}$, $r = k_1 M^2$, and $c_{p,K}$ is a constant depending only on $\mathcal{X}$, p, K.

Proof. Denote

$$\mathcal{F}_r = \{g(z) = (y - \pi(f)(x))^2 - (y - f_\rho(x))^2 : f \in \mathcal{B}_r\}.$$

We can see that $Eg = \mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho)$ and $\frac{1}{n}\sum_{i=1}^n g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f_\rho)$. Since $\|\pi(f)\|_\infty \leqslant M$ and $|f_\rho(x)| \leqslant M$, we have

$$|g(z)| = \left| \left(\pi(f)(x) - f_\rho(x) \right) \left((\pi(f)(x) - y) + (f_\rho(x) - y) \right) \right| \\ \leqslant 8M^2$$

and

$$Eg^2 = \int_{\mathcal{Z}} (\pi(f)(x) - f_\rho(x))^2 ((\pi(f)(x) - y) \\ + (f_\rho(x) - y))^2 d\rho \leqslant 16M^2 Eg.$$

For $g_1, g_2 \in \mathcal{F}_r$, we have

$$|g_1(z) - g_2(z)| = |(y - \pi(f_1)(x))^2 - (y - \pi(f_2)(x))^2| \\ \leqslant 4M|\pi(f_1)(x) - \pi(f_2)(x)| \\ \leqslant 4M|f_1(x) - f_2(x)|.$$

Then

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_r, \epsilon) \leqslant \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_r, \frac{\epsilon}{4M} \right) \leqslant \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_1, \frac{\epsilon}{4Mr} \right).$$

In connection with Definition 6 and Theorem 2 in Shi et al. (2011), this implies

$$\log \mathcal{N}_2\left(\mathcal{B}_1, \frac{\epsilon}{4Mr} \right) \leqslant c_{p,K}(4Mr)^p \epsilon^{-p}, \quad \forall \epsilon > 0,$$

where $c_{p,K}$ is a constant independent of ϵ.

Applying Lemma 2 with $B = c = 16M^2$ and $a = c_{p,K}(4Mr)^p$, for any $\delta \in (0, 1)$ and $\forall g \in \mathcal{F}_r$,

$$Eg - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \leqslant \frac{1}{2} Eg + c'_p (16M^2)^{\frac{2-p}{2+p}} \left(\frac{c_{p,K}(4Mr)^p}{n} \right)^{\frac{2}{2+p}} \\ + 320M^2 \frac{\log(1/\delta)}{n}, \\ \leqslant \frac{1}{2} Eg + 640M^2 (c_{p,K}(4Mr)^p)^{\frac{2}{2+p}} \\ \times \log(1/\delta) n^{-\frac{2}{2+p}}$$

holds with the confidence $1-\delta$. Note that $\hat{f}_k \in \mathcal{B}_r$. Then $\{\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi(\hat{f}_k)) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}$ can be bounded by $\frac{1}{2}\{\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho)\} + c_1 \log(1/\delta) n^{-\frac{2}{2+p}}$. This completes the proof. □

4.2. *Estimation of the hypothesis error*

A good estimation of the hypothesis error is important to achieve tight generalization error bounds for learning with the data-dependent hypothesis spaces. In Shi et al. (2011) and Wu and Zhou (2008), the hypothesis error has been well studied for the regularized method with data-dependent hypothesis spaces. Different from these studies, we establish the estimation of the hypothesis error $\mathcal{E}_{\mathbf{z}}(\hat{f}_k) - \mathcal{E}_{\mathbf{z}}(f_\lambda)$ based on Theorem 2.3 in Barron et al. (2008).

Proposition 3. *For any $\delta > 0$ and $k \geqslant T$, the inequality*

$$\mathcal{E}_{\mathbf{z}}(\hat{f}_k) - \mathcal{E}_{\mathbf{z}}(f_\lambda) \leqslant \min\left\{k_1^2 k_0^{-2}, \left(1 + k_1 k_0^{-1}\sqrt{\frac{\log(n/\delta)}{2n}}\right)^2\right\} \times \frac{D^2(\lambda)}{k\lambda^2}$$

is true with the confidence at least $1-\delta$.

Proof. Denote the function space

$$\mathcal{H}_1^n = \left\{h = \sum_i \alpha_i^n \tilde{K}_{u_i}^n : \alpha_i^n = \alpha_i \|\tilde{K}_{u_i}\|_n, \quad \tilde{K}_{u_i}^n = \tilde{K}_{u_i}/\|\tilde{K}_{u_i}\|_n, \sum_i \alpha_i \tilde{K}_{u_i} \in \mathcal{H}_1\right\}$$

with the norm

$$\|f\|_{\mathcal{H}_1^n} = \inf\left\{\sum_i |\alpha_i^n| : f = \sum_i \alpha_i \tilde{K}_{u_i}\right\}.$$

From Theorem 2.3 and the inequality (3.26) in Barron et al. (2008), we know

$$\mathcal{E}_{\mathbf{z}}(\hat{f}_k) - \mathcal{E}_{\mathbf{z}}(f_\lambda) \leqslant \frac{4\|f_\lambda\|^2_{\mathcal{H}_1^n}}{k}. \tag{4.3}$$

Since $\|f_\lambda\|^2_{\mathcal{H}_1^n}$ depends on $\mathbf{z}$, we must further find its relation with $\|f_\lambda\|^2_{\mathcal{H}_1}$. From the definitions of $\|f\|_{\mathcal{H}_1^n}$ and $\|f\|_{\mathcal{H}_1}$, we know that $\|f\|_{\mathcal{H}_1^n} \leqslant \frac{k_1}{k_0}\|f\|_{\mathcal{H}_1}$.

We also observe that $\|\tilde{K}_{u_i}\|^2_{L^2_{\rho_{\mathcal{X}}}} = E\tilde{K}^2_{u_i} = 1$ and

$$\|\tilde{K}_{u_i}\|_n - 1 = \sqrt{\frac{1}{n}\sum_{j=1}^n |\tilde{K}(u_i, x_j)|^2} - 1 \leqslant \frac{1}{n}\sum_{j=1}^n |\tilde{K}(u_i, x_j)|^2 - 1. \tag{4.4}$$

Meanwhile, based on the Hoeffding inequality, for any i, we have

$$\mathrm{Prob}\left\{\frac{1}{n}\sum_{j=1}^n |\tilde{K}(u_i, x_j)|^2 - E\tilde{K}^2_{u_i} \geqslant \epsilon\right\} \leqslant \exp\left\{-\frac{2k_0^2\epsilon^2 n}{k_1^2}\right\}. \tag{4.5}$$

By setting $\delta = \exp\{-\frac{2k_0^2\epsilon^2 n}{k_1^2}\}$, from (4.4) and (4.5), we have with the confidence $1-\delta$,

$$\|\tilde{K}_{u_i}\|_n \leqslant \frac{1}{n}\sum_{j=1}^n |\tilde{K}(u_i, x_j)|^2 \leqslant E\tilde{K}_{u_i} + k_1 k_0^{-1}\sqrt{\frac{\log(1/\delta)}{2n}} \leqslant 1 + k_1 k_0^{-1}\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Hence,

$$\|f_\lambda\|^2_{\mathcal{H}_1^n} \leqslant \left(1 + k_1 k_0^{-1}\sqrt{\frac{\log(1/\delta)}{2n}}\right)^2 \|f_\lambda\|^2_{\mathcal{H}_1}$$

is true with the confidence at least $1 - n\delta$.

Finally, combining the above inequality with (4.3) and $\|f_\lambda\|_{\mathcal{H}_1} \leqslant D(\lambda)/\lambda$, we derive the desired result. □

4.3. *Estimate of learning rates*

Based on the above estimations of the sample and hypothesis errors, we derive the estimation of the learning rates.

Proof of Theorem 1. Combining the results in Propositions 1–3, we have that

$$\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho) \leqslant \frac{14(3M + k_1 D(\lambda)/\lambda)}{3n}\log(1/\delta) + 3D(\lambda) + 2c_1 \log(1/\delta) n^{-\frac{2}{2+p}} + 2\min\left\{k_1 k_0^{-1}, \left(1 + k_1 k_0^{-1}\sqrt{\frac{\log(n/\delta)}{2n}}\right)^2\right\} \times \frac{D^2(\lambda)}{k\lambda^2}$$

with the confidence at least $1-3\delta$. From the condition of $D(\lambda)$, we have with confidence $1-\delta$

$$\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho) \leqslant c\log(1/\delta)\left(\frac{\lambda^{2q-2}}{n} + n^{-\frac{2}{2+p}} + \lambda^q + \frac{\lambda^{2q-2}}{k}\right),$$

where c is a constant independent of n, δ, k. Note that $k \geqslant T \geqslant n$. Then, setting $\lambda = n^{\frac{1}{q-2}}$, we derive

$$\mathcal{E}(\pi(\hat{f}_k)) - \mathcal{E}(f_\rho) \leqslant \tilde{c}\log(1/\delta) n^{-\min\left\{\frac{2}{2+p}, \frac{q}{2-q}\right\}}$$

with the confidence $1-\delta$. This finishes the proof. □

5. An empirical study

Our theoretical analysis of the semi-supervised greedy algorithm (SSG) shows that it is efficient to achieve fast learning rates for the regression learning. In this section, we compare our method with the least square regularized regression (LSR) algorithm in RKHS.

The least square regularized regression algorithm has been extensively studied in learning theory (Cucker & Zhou, 2007) and can be formulated as

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_K}\left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}.$$

We consider $\mathcal{X} = [0, 1]$, the Gaussian kernel $K(x, t) = \exp(-\frac{(x-t)^2}{2\mu^2})$ with $\mu = 1$, and $\lambda = 10^{-3}$. We choose $f_\rho(x) = \sin(\pi x)$ and $f_\rho(x) = x^2$ as the target functions respectively. The labeled samples (x_i, y_i) are generated as follows: x_i is independently and uniformly distributed within $[0, 1]$ and

$$y_i = f_\rho(x_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma_\epsilon^2).$$

Three different noise levels are considered: $\sigma_\epsilon = 0.01, 0.05, 0.25$. The unlabeled samples $\mathbf{x} = \{\tilde{x}_j\}_{j=1}^{1000}$ are also independently and uniformly distributed within $[0, 1]$. The mean square error (MSE) of f on $\mathbf{x}$ is defined as below

$$\mathrm{MSE} = \frac{1}{1000}\sum_{j=1}^{1000}(f(\tilde{x}_j) - f_\rho(\tilde{x}_j))^2,$$

which is used to measure the efficiency of learning algorithms.

We report the mean value of the MSE results of the 100 repeating tests for each case. The results are summarized in Figs. 5.1 and 5.2.

The results show that the prediction performance depends on not only the smoothness of the target function but also the noise level, and that the SSG might be better for the small noise situation. This preliminary study shows that our method is efficient for regression.

Finally, we remark that the upper bound analysis and preliminary empirical analysis are not enough for a comprehensive theoretical understanding of the proposed method. The lower bound analysis is also important to evaluate its learning performance. For

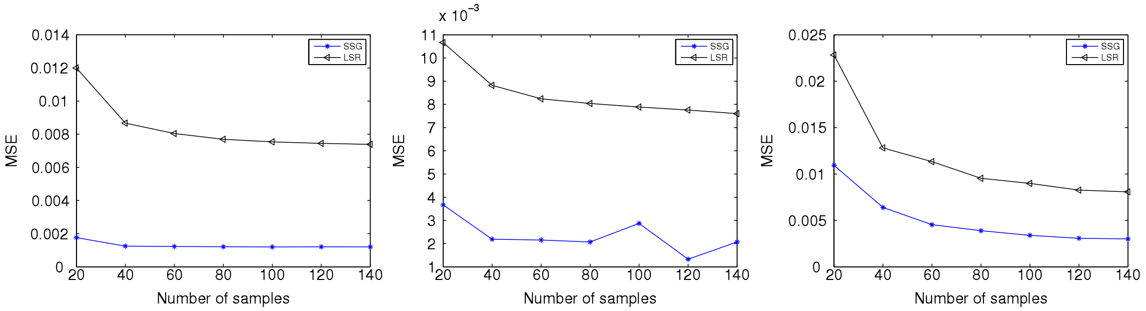


Fig. 5.1. Empirical comparison of $g(x) = \sin(\pi x)$ and $\sigma_\epsilon = 0.01, 0.05, 0.25$.

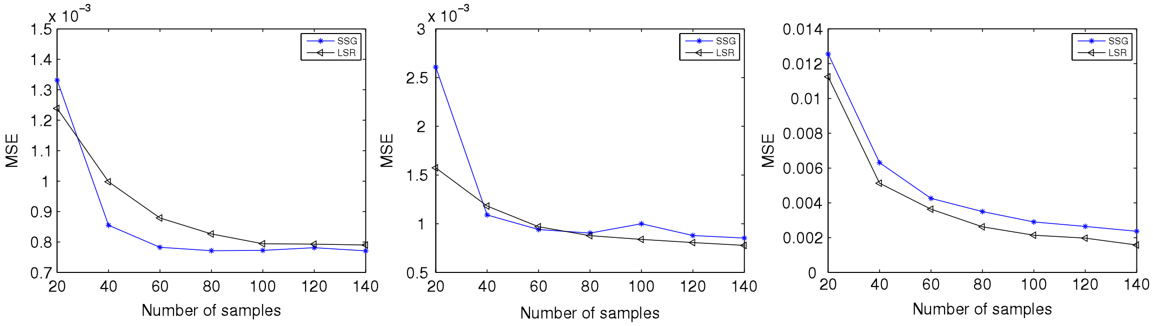


Fig. 5.2. Empirical comparison of $g(x) = x^2$ and $\sigma_\epsilon = 0.01, 0.05, 0.25$.

the semi-supervised learning, it is crucial to utilize the unlabeled data to find the characteristics of the target function under a certain assumption of the distribution. They are out of the scope of this paper and we leave it for our future study.

6. Conclusion and discussion

This paper has introduced a sparse semi-supervised method to learn the regression functions from samples using the orthogonal greedy algorithm. Fast learning rates were derived under mild assumptions. The symmetric or positive semi-definite demand for kernel and the interior cone condition for $\mathcal{X}$ (see Shi et al., 2011) is abandoned in this paper. There are some extensions to this method which we discuss as below:

1. The semi-supervised learning based on other greedy algorithms: The proposed method depends on the orthogonal greedy algorithm. It remains open to explore the semi-supervised learning with other greedy algorithms, e.g., the pure, relaxed, or stepwise greedy algorithm. Some ideas are presented in Barron et al. (2008) and Zhang (2002) for the supervised settings.
2. Sample error estimation with different techniques: Our analysis shows that the work for the unlabeled data is limited. Under current analysis techniques, the sample error cannot be improved even if the number of the unlabeled data tends to infinity. For this reason, it would be important to explore a new technique to analyze the sample error under suitable assumptions.

Acknowledgments

The authors would like to thank the reviewers for their valued comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 11001092, No. 11071058, and No. 11226304; by the Fundamental Research Funds for the Central Universities (Program No. 2011PY130), and by the Macau Science and Technology Development Fund (FDCT) under grant 017/2012/A1 and the Research Committee at University of Macau under grants MYRG113(Y1-L3)-FST12-ZYC, SRG010-FST11-TYY, MYRG187(Y1-L3)-FST11-TYY, MYRG205(Y1-L4)-FST11-TYY, and MRG001/ZYC/2013/FST.

References

Barron, A. R., Cohen, A., Dahmen, W., & DeVore, R. (2008). Approximation and learning by greedy algorithm. *Annals of Statistics*, *36*, 64–94.

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, *56*, 209–239.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularizaion: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*, 2399–2434.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *11th annual conference on computational learning theory* (pp. 92–100).

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge, Massachusetts: MIT Press.

Chen, H., & Li, L. Q. (2009). Semi-supervised multi-category classification with imperfect model. *IEEE Transactions on Neural Networks*, *20*, 1594–1603.

Chen, H., Li, L. Q., & Peng, J. T. (2009). Error bounds of semi-supervised multi-graph regularized classifiers. *Information Sciences*, *179*, 1960–1969.

Chen, H., Li, L. Q., & Peng, J. T. (2010). Semi-supervised learning based on high density regions estimation. *Neural Networks*, *23*, 812–818.

Chen, D. R., Wu, Q., Ying, Y., & Zhou, D. X. (2004). Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, *5*, 1143–1175.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

Cucker, F., & Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*. Cambridge, UK: Cambridge Univ. Press.

Fan, M., Gu, N., Qiao, H., & Zhang, B. (2011). Sparse regualrization for semi-supervised classification. *Pattern Recognition*, *44*, 1777–1784.

Johnson, R., & Zhang, T. (2007). On the effectiveness of Laplacian normalization for graph-based semi-supervised learning. *Journal of Machine Learning Research*, *8*, 1489–1517.

Johnson, R., & Zhang, T. (2008). Graph-based semi-supervised learning and spectral kernel design. *IEEE Transactions on Information Theory*, *54*, 275–288.

Nair, P. B., Choudhury, A., & Keane, A. J. (2007). Some greedy learning algorithms for sparse regression and classification with Mercer kernels. *Journal of Machine Learning Research*, *3*, 781–801.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, *8*, 1369–1392.

Shi, L., Feng, Y. L., & Zhou, D. X. (2011). Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, *31*, 286–302.

Sindhwani, V., Niyogi, P., & Belkin, M. (2005). A co-regularization approach to semi-supervised learning with multiple views. In *Proceedings of the workshop on learning with multiple views, 22nd ICML*.

Sindhwani, V., & Rosenberg, D. (2008). An RKHS for multi-view learning and manifold co-regularization. In *Proceeding of the 25th ICML* (pp. 976–983).

Sun, S., & Shawe Taylor, J. (2010). Sparse semi-supervised learning using conjugate functions. *Journal of Machine Learning Research*, *11*, 2423–2455.

Sun, H. W., & Wu, Q. (2011). Least square regression with indefinite kernels and coefficient regularization. *Applied and Computational Harmonic Analysis*, *30*, 96–109.

Tsang, I. W., & Kwok, J. T. (2007). Large-scale sparsified manifold regularization. In *Advances in NIPS 19* (pp. 1401–1408).

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

Wu, Q., & Zhou, D. X. (2008). Learning with sample dependent hypothesis spaces. *Computers & Mathematics with Applications*, *56*, 2896–2907.

Xiao, Q. W., & Zhou, D. X. (2010). Learning by nonsymmetric kernel with data dependent spaces and ℓ^1-regularizer. *Taiwanese Journal of Mathematics*, *14*, 1821–1836.

Zhang, T. (2002). Approximation bounds for some sparse kernel regression algorithms. *Neural Computation*, *14*, 3013–3042.

Zhang, T. (2009). On the consistency of feature selection using greedy least squres regression. *Journal of Machine Learning Research*, *10*, 555–568.

Zhu, X. (2005). Semi-supervised learning literature survey. Technical report 1530. Computer Sciences. University of Wisconsin–Madison.

Zou, B., Li, L. Q., & Xu, Z. B. (2009). The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning*, *75*, 275–295.