J. Math. Anal. Appl. 386 (2012) 205–212

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Semi-Supervised Learning with the help of Parzen Windows

Shao-Gao Lv [a,*], Yun-Long Feng [b]

[a] *Statistics School, Southwestern University of Finance and Economics, Chengdu 611130, China*
[b] *Joint Advanced Research Center of University of Science and Technology of China, and City University of Hong Kong, SuZhou 215123, China*

ARTICLE INFO

Article history:
Received 5 April 2011
Available online 29 July 2011
Submitted by U. Stadtmueller

Keywords:
Semi-Supervised Learning
Graph-based models
Support vector machine
Least square regression
Reproducing kernel Hilbert spaces

ABSTRACT

Semi-Supervised Learning is a family of machine learning techniques that make use of both labeled and unlabeled data for training, typically a small amount of labeled data with a large number of unlabeled data. In this paper we propose a Semi-Supervised regression algorithm by means of density estimator, generated by Parzen Windows functions under the framework of Semi-Supervised Learning. We conduct error analysis by capacity independent technique and obtain some satisfactory learning rates in terms of regularity of the target function and the decay condition on the marginal distribution near the boundary.

1. Introduction

Traditional learning from example uses only labeled data to train. However, in many real world application, it is relatively easy to acquire a large amount of unlabeled data. For example, in phonological acquisition contexts, a child is exposed to many acoustic utterances. These utterances do not come with identifiable phonological markers. Corrective feedback is the main source of directly labeled examples. In many cases, a small amount of feedback is sufficient to allow the child to master the acoustic-to-phonetic mapping of any language. Another instance is for a robot with a video camera. The robot continuously takes high frame-rate video of its surroundings and seek to learn the names of various objects in the video, but the robot stores names in designed system beforehand only very rarely. These phenomenon can be formulated as a Semi-Supervised Learning situation: most objects are unlabeled, while only a few are labeled. Semi-Supervised Learning addresses this problem by using large amount of unlabeled data, together with the labeled data, to build better learners. Due to less human efforts and higher accuracy, Semi-Supervised Learning can be of great practical value.

At the first glance, the unlabeled data by itself does not carry any information on the mapping from the input space X to output space Y. How can it help us learn a better predictor $f : X \to Y$. Balcan and Blum pointed out in [1] that the key lies in an implicit ordering of $f \in \mathcal{F}$ induced by the unlabeled data. Roughly speaking, if the implicit order happens to rank the target predictor f^* near the top, then one needs less labeled data to learn f^*. This idea will be formalized on using PAC learning bounds. In other context, the implicit ordering is interpreted as a prior over $\mathcal{F}$ or as a regularizer. Therefore, finding the implicit order of input space and designing a good learning algorithm are both very critical in Semi-Supervised Learning.

The history of Semi-Supervised Learning goes back to at least the 70s, when Self-training, transduction and Gaussian mixtures with the EM algorithm first emerged. A variety of Semi-Supervised techniques have been developed, which all can be classified into generative and discriminative methods. A straightforward, generative Semi-Supervised method is the Expectation–Maximization (EM) algorithm. The EM approach for naive Bayes text classification models is discussed

[*] Corresponding author.
E-mail address: kenan716@mail.ustc.edu.cn (S.-G. Lv).

0022-247X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmaa.2011.07.059

by Nigam [12]. Discriminative Semi-Supervised methods [9] include probabilistic and non-probabilistic approaches, such as transductive or Semi-Supervised Support vector Machines (TSVMs, S3VMs) and a variety of other graph based methods [23,2]. Different learning methods for Semi-Supervised Learning are based on different assumptions. Fulfilling these assumptions is crucial for the success of the methods. More discussions can be founded in [3,22].

From the learning kernel point of view, many existing structured learning algorithms (e.g. conditional random fields, maximum margin Markov networks) can be endowed with a Semi-Supervised kernel. The key idea lies in that the graph kernel can be constructed with all the unlabeled data, next one applies the graph kernel to a standard structured learning kernel machine. Such kernel machines include the kernelized conditional random fields [10] and the maximum margin Markov networks [17], which differ primarily by the loss function they use.

As is well known, Parzen Windows method [13] is a widely used technique for kernel density estimation and regression (see e.g. [19]), and recently they are also applied to multi-classifier learning [14] and regression problem [21] in a general subset of $\mathbb{R}^d$ by a decay condition near the boundary.

These motivate us to study Semi-Supervised regression integrated with density estimations using Parzen Windows. Intuitively, the unknown marginal distribution can be characterized well by Parzen Windows using amounts of unlabeled data. This amounts to provide us some additional useful information for supervised learning.

Definition 1. We say that $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a basic window function if it is continuous and it satisfies:

(i) $\int_{\mathbb{R}^d} \Phi(x, y)\, dy = 1$ for each $x \in \mathbb{R}^d$;
(ii) there exist some $q > d + 2$ and constant $c_q > 0$ such that

$$\left|\Phi(x, y)\right| \leqslant \frac{c_q}{(1 + |x - y|)^q}, \quad \forall x, y \in \mathbb{R}^d. \tag{1.1}$$

The decay condition (1.1) is reasonable since it is satisfied by many commonly used function in multivariate analysis and learning theory.

Suppose that the input space X is a compact subset of $\mathbb{R}^d$, and the output space $Y \subseteq \mathbb{R}$. Let ρ be the probability measure defined on $X \times Y$. Given a sequence of sample $\mathbf{z} := \{x_i, y_i\}_{i=1}^m$ drawn independently according to ρ, and n unlabeled data $\{x_i\}_{i=m+1}^{m+n}$ drawn from the marginal distribution denoted by ρ_X, often $m \ll n$. In our setting, the objective is to learn the regression function

$$f_\rho(x) = \int_Y y\, d\rho(y|x), \quad x \in X,$$

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

We consider a learning algorithm in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with a Mercer kernel K, where $K : X \times X \to \mathbb{R}$ is a continuous, symmetric and positive semi-definite function. Without loss of generality, assume that $\sup_{x,y \in X} |K(x, y)| \leqslant \kappa$.

On the basis of density estimator generated by Parzen Windows, we learn the regression function by

$$f_{\mathbf{z},\lambda}^{\sigma} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m(m+n)\sigma^d} \sum_{i=1}^{m} \sum_{j=1}^{m+n} \Phi\left(\frac{x_i}{\sigma}, \frac{x_j}{\sigma}\right)\left(f(x_i) - y_i\right)^2 + \lambda \|f\|_K^2 \right\}, \tag{1.2}$$

where $\sigma = \sigma(m) > 0$ is a scalar parameter and $\lambda \in (0, 1)$, a regularization parameter controlling the trade-off between the empirical error and the penalty term $\|f\|_K^2$.

Since X is a compact subset of $\mathbb{R}^d$, some regularity conditions on both the marginal distribution and the density are required. Throughout the paper, we give the following assumption.

Assumption 1. Suppose that for some $0 < \tau \leqslant 1$, $c_p > 0$, the marginal distribution ρ_X satisfies

$$\rho_X\left(\left\{x \in X\colon \inf_{u \in R^d \setminus X} |u - x| \leqslant s\right\}\right) \leqslant c_p^2 s^{2\tau}, \quad \forall s > 0, \tag{1.3}$$

and the density $p(x)$ of ρ_X satisfies

$$\sup_{x \in X} p(x) \leqslant c_p, \qquad \left|p(x) - p(v)\right| \leqslant c_p |v - x|^{\tau}, \quad v, x \in X. \tag{1.4}$$

Denote L_μ^2 as the L^2-space with the inner production $\langle f, g\rangle_{L_\mu^2} = \int_X f(x)g(x)\, d\mu$. Our learning rates will be achieved under the regularity assumption on the regression function that f_ρ lies in the range of L_K^r for some $r > 0$. Here L_K is the integral operator on L_ν^2 defined by

$$L_K(f)(x)=\int\limits_X K(x,y)f(y)\,d\nu(y),\quad x\in X,\ f\in L^2_\nu,$$

where $d\nu(x)=\frac{p(x)}{V_p}d\rho_X(x)$ with $V_p:=\int_X p(x)^2 d(x)>0$. The operator L_K is linear, compact, positive and can be also regarded as a self-adjoint operator on $\mathcal{H}_K$. Since the measure $d\nu(x)=\frac{p(x)}{V_p}d\rho_X(x)$ is probability one on X and $\|p\|_\infty\leqslant c_p$, we see that the fractional operator L_K^r $(0<r\leqslant 1)$ is also well-defined on $L^2_{\rho_X}$ in [4].

We can now state our learning rate for the Semi-Supervised regression algorithm which will be proved in Section 5.

Theorem 1. *Assume $|y|\leqslant M$ for some constant $M>0$. Suppose Assumption 1 holds, and $L_K^{-r}f_\rho\in L^2_{\rho_X}$ with $0<r\leqslant 1$. Choose*

$$\lambda=\sigma^\tau\quad and\quad \lambda=\left(\frac{1}{m}\right)^{\frac{\tau}{2\tau(r+1)+d}}.$$

For any $0<\delta<1$, with confidence to $1-\delta$, there holds

$$\|f^\sigma_{\mathbf{z},\lambda}-f_\rho\|_{L^2_{\rho_X}}\leqslant\widetilde{C}\log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\frac{\tau r}{2\tau(r+1)+d}},$$

where $\widetilde{C}$ is a constant independent on m or δ.

The key technique behind the proof for the convergence of the learning scheme lies in the error decomposition, consisting of a sample error term and an approximation error term. The first term, the sample error, is bounded using a concentration inequality in a Hilbert space since it is a function of the sample **z**. On the other hand, the second term, the approximation error, does not depend on the sample and we use approximation theory to attain the purpose.

The paper is organized as follows. In Section 2 a new Semi-Supervised regression algorithm is proposed based on density estimations. Then we introduce a necessary concentration inequality in a Hilbert space and sample error in $\mathcal{H}_K$ are estimated in Section 3. In Section 4 we bound the approximation error under the condition of Assumption 1 and the regularity of the regression function. Finally we obtain the learning rate by combining the sample error with approximation error.

2. Semi-Supervised algorithm with density estimators

Turn our attention to Parzen Windows again. Let $(x_1,x_2,\ldots,x_n)$ be an i.i.d. sample drawn from some distribution with an unknown density p. We are interested in estimating the shape of this function p. Its kernel density estimator by means of Parzen Windows method is given as

$$p_{\mathbf{x},\sigma}=\frac{1}{n\sigma^d}\sum_{i=1}^n\Phi\left(\frac{x}{\sigma},\frac{x_i}{\sigma}\right),\tag{2.1}$$

where $\sigma>0$ is a smoothing parameter called the bandwidth. Parzen showed that $p_{\mathbf{x},\sigma}$ converges to $p(x)$ and $\sigma=\sigma(n)$ satisfies

$$\lim_{n\to\infty}\sigma(n)=0,\qquad \lim_{n\to\infty}n\big[\sigma(n)\big]^d=\infty.$$

A kernel with subscript σ is called the scaled kernel and defined as $\Phi_\sigma(x)=1/\sigma\Phi(x/\sigma)$. Intuitively one wants to choose σ as small as the data allows, however there is always a trade-off between the bias of the estimator and its variance. When the boundary of X satisfies certain decay condition, by choosing a proper parameter $\sigma(n)$, a standard convergence rate for density estimation founded in [6,8] can be expressed as

$$\|p_{\mathbf{x},\sigma}-p(x)\|_{L^2(X)}=\mathcal{O}\big(n^{-2/(d+4)}\big).\tag{2.2}$$

Note that if the unknown density p is smooth enough, the convergence rate can be improved by using the higher order Parzen Windows methods [21]. In this paper, our idea lies in that more accuracy should be given at sample point where the density of the marginal distribution ρ_X is much larger than other field. In the classical regression algorithm, all the sample points are equivalently treated and it may cause large violation when there exist some singular points among all the sample data. Since an unseen data will appears with greater confidence in much denser fields, our efforts should focus on there by using huge surplus unlabeled data. Considering the data structure of the input space, we propose the kernel-based regression estimator with respect to density function as

$$f_{\mathbf{z},\lambda}=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^m p(x_i)\big(f(x_i)-y_i\big)^2+\lambda\|f\|_K^2\right\}.$$

As mentioned above, $p(x)$ is usually unknown and we shall replace it with a density estimator. Here we make use of Parzen Windows methods as the kernel density estimation. Thus, given a set of n unlabeled examples $\{x_i\}_{i=m+1}^{m+n}$, we get the optimization problem (1.2).

The optimization method (1.2) looks like weighted least square regression (WLSR) in Supervised Learning [5,7]. But here we are mainly concerned with underling data shape induced by large amounts of unlabeled data. In the classical WLSR setting, there exists no surplus unlabeled data and estimation parameter may be sensitive to the weight function using only a few observations.

By the reproducing property of RKHS, the Representer Theorem holds, which means that the minimizer can be obtained in a finite dimensional space in terms of both labeled and unlabeled examples. As before, the Representer Theorem shows that the solution has the form

$$f_{\mathbf{z},\lambda}^{\sigma}(x)=\sum_{i=1}^{m}\alpha_i^{*}K(x,x_i).$$

Substituting this form in the problem above, we arrive at the following convex differentiable objective function of the m-dimensional variable $a=[a_1,\ldots,a_m]^T$:

$$a^{*}=\arg\min\frac{1}{m}(Y-K_{\mathbf{x}}a)^{T}Q(Y-K_{\mathbf{x}}a)+\lambda a^{T}K_{\mathbf{x}}a,$$

where $K_{\mathbf{x}}$ is the $m\times m$ gram matrix $K_{i,j}=K(x_i,x_j)$, $Q=\operatorname{diag}(p_{\mathbf{x},\sigma}(x_1),\ldots,p_{\mathbf{x},\sigma}(x_m))$ and Y is the label vector $Y=[y_1,\ldots,y_m]^T$. The derivative of the objective function vanishes at the minimizer, which leads to the following solution:

$$\alpha^{*}=(QK_{\mathbf{x}}+\lambda mI)^{-1}QY.$$

3. Sample error estimate

In this section we investigate the approximation error of our proposed algorithm as the parameters σ and λ change.

To understand (1.2), denote by $\mathbf{x}$ the set of inputs $\{x_1,\ldots,x_m\}$, and define the sampling operator $\mathbf{S_x}:\mathcal{H}_K\to\mathbb{R}^m$ as $\mathbf{S_x}(f)=(f(x_i))_{i=1}^m$. The adjoint of the sampling operator, $\mathbf{S_x^T}:\mathbb{R}^m\to\mathcal{H}_K$, can be written as

$$\mathbf{S_x^T}c=\sum_{i=1}^{m}c_iK_{x_i},\quad c\in\mathbb{R}^m.$$

A similar methods with [4] are applied to this algorithm (1.2), whose minimizer can be represented by

$$f_{\mathbf{z},\lambda}^{\sigma}=\left(\frac{1}{m(m+n)\sigma^d}\mathbf{S_x^T}G\mathbf{S_x}+\lambda I\right)^{-1}\frac{1}{m(m+n)\sigma^d}\mathbf{S_x^T}GY, \tag{3.1}$$

where $G=\operatorname{diag}\{g(x_1),\ldots,g(x_m)\}$ with $g(x_i)=\sum_{j=1}^{m+n}\Phi(\frac{x_i}{\sigma},\frac{x_j}{\sigma})$ $(i=1,\ldots,m)$ and Y is defined as above.

To study the limitation behavior of the function $f_{\mathbf{z},\lambda}^{\sigma}$, the following notations play key role in our theoretical analysis.

Definition 2. The regularization error and the regularizing function of the triple $(\mathcal{H}_K,f_\rho,\rho_X)$ are defined respectively as

$$\mathcal{D}(\lambda)=\inf_{f\in\mathcal{H}_K}\left\{\frac{1}{\sigma^d}\int_X\int_X\Phi\left(\frac{x}{\sigma},\frac{t}{\sigma}\right)\left(f(x)-f_\rho(x)\right)^2d\rho_X(x)\,d\rho_X(t)+\lambda\|f\|_K^2\right\}, \tag{3.2}$$

and

$$f_\lambda^\sigma=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{\sigma^d}\int_X\int_X\Phi\left(\frac{x}{\sigma},\frac{t}{\sigma}\right)\left(f(x)-f_\rho(x)\right)^2d\rho_X(x)\,d\rho_X(t)+\lambda\|f\|_K^2\right\},\quad\lambda>0. \tag{3.3}$$

In order to obtain the explicit form of f_λ^σ, we introduce the following definition $L_K^\sigma:L_{\rho_X}^2\to L_{\rho_X}^2$, given as

$$\left(L_K^\sigma f\right)(u)=\frac{1}{\sigma^d}\int_X\int_X\Phi\left(\frac{x}{\sigma},\frac{t}{\sigma}\right)f(x)K_u(x)\,d\rho_X(x)\,d\rho_X(t),\quad\text{for }\forall u\in X.$$

Taking the functional derivatives, we know that f_λ^σ can be expressed associated with L_K^σ

$$f_\lambda^\sigma=\left(\lambda I+L_K^\sigma\right)^{-1}L_K^\sigma f_\rho. \tag{3.4}$$

Intuitively, fixed $f \in \mathcal{H}_K$, there holds $\mathbb{E}(\frac{1}{l(l+u)\sigma^d}\mathbf{S}_\mathbf{x}^\mathbf{T} G \mathbf{S}_\mathbf{x})f = (L_K^\sigma)f$, and $\mathbb{E}(\frac{1}{l(l+u)\sigma^d}\mathbf{S}_\mathbf{x}^\mathbf{T} GY) = L_K^\sigma f_\rho$. Therefore, it seems that $f_{\mathbf{x},\sigma}$ can approximate f_λ^σ.

To bound the sample error $\|f_{\mathbf{z},\lambda}^\sigma - f_\lambda^\sigma\|$, we need to introduce a McDiarmid–Bernstein type probability inequality for vector-valued random variables, which appears in [15].

Lemma 1. *Let $\mathbf{x} = \{x_i\}_{i=1}^m$ be i.i.d. draws from a probability distribution ρ on X, $(\mathcal{H}, \|\cdot\|)$ be a Hilbert space, and $F : X^m \to \mathcal{H}$ be measurable. If there exists $\widetilde{M} \geqslant 0$ such that $\|F(\mathbf{x}) - E_{x_i}(F(\mathbf{x}))\| \leqslant \widetilde{M}$ for each $1 \leqslant i \leqslant m$ and almost $\mathbf{x} \in X^m$, then for every $\varepsilon > 0$,*

$$Prob_{\mathbf{x}\in X^m}\left\{\left\|F(\mathbf{x}) - E_\mathbf{x}\big(F(\mathbf{x})\big)\right\| \geqslant \varepsilon\right\} \leqslant 2\exp\left\{-\frac{\varepsilon^2}{2(\widetilde{M}\varepsilon + \widetilde{\sigma}^2)}\right\}, \tag{3.5}$$

where $\widetilde{\sigma}^2 := \sum_{i=1}^m \sup_{\mathbf{x}\setminus x_i \in X^{m-1}} E_{x_i}\{\|F(\mathbf{x}) - E_{x_i}(F(\mathbf{x}))\|^2\}$. For any $0 < \delta < 1$, with confidence to $1 - \delta$, there holds

$$\left\|F(\mathbf{x}) - E_\mathbf{x}\big(F(\mathbf{x})\big)\right\| \leqslant 2\log\frac{2}{\delta}\left\{\widetilde{M} + \sqrt{\widetilde{\sigma}^2}\right\}.$$

It is worth pointing out that though estimating the Sample Error is a standard method in learning theory [11,16], the operator L_K^σ is not necessary in the Hilbert–Schmidt space and the corresponding skills are no longer applicable.

Proposition 1. *Suppose that the density $p(x)$ of ρ_X exists and satisfies $\sup_{x\in X} p(x) \leqslant c_p$, then for any $0 < \sigma \leqslant 1$ and $0 < \delta \leqslant 1$, with confidence $1 - \delta$, there holds*

$$\left\|f_{\mathbf{z},\lambda}^\sigma - f_\lambda^\sigma\right\|_K \leqslant \log\left(\frac{2}{\delta}\right)\frac{12c_q\kappa(M + \|f_\lambda^\sigma\|_K)}{\sqrt{m}\lambda\sigma^{d/2}}\left(\frac{1}{\sqrt{m}\sigma^{d/2}} + c_q c_p\right). \tag{3.6}$$

Proof. By (3.1), we have

$$f_{\mathbf{z},\lambda}^\sigma - f_\lambda^\sigma = \left(\frac{1}{m(m+n)\sigma^d}\mathbf{S}_\mathbf{x}^\mathbf{T} G\mathbf{S}_\mathbf{x} + \lambda I\right)^{-1}\left\{\frac{1}{m(m+n)\sigma^d}\big(\mathbf{S}_\mathbf{x}^\mathbf{T} GY - \mathbf{S}_\mathbf{x}^\mathbf{T} G\mathbf{S}_\mathbf{x} f_\lambda^\sigma\big) - \lambda f_\lambda^\sigma\right\}.$$

Define a function $F : Z^m \times X^n \to \mathcal{H}_K$ by

$$\begin{aligned} F(\mathbf{z}, \mathbf{x}) &= \frac{1}{m(m+n)\sigma^d}\big(\mathbf{S}_\mathbf{x}^\mathbf{T} GY - \mathbf{S}_\mathbf{x}^\mathbf{T} G\mathbf{S}_\mathbf{x} f_\lambda^\sigma\big) \\ &= \frac{1}{m(m+n)\sigma^d}\left\{\sum_{i=1}^m\sum_{j=1}^{m+n}\Phi\left(\frac{x_i}{\sigma}, \frac{x_j}{\sigma}\right)y_i K_{x_i} - \sum_{i=1}^m\sum_{j=1}^{m+n}\Phi\left(\frac{x_i}{\sigma}, \frac{x_j}{\sigma}\right)K_{x_i} f_\lambda^\sigma(x_i)\right\}. \end{aligned}$$

By independence, the expectation of $F(\mathbf{z}, \mathbf{x})$ equals

$$\frac{1}{\sigma^d}\left\{\int_X\int_X \Phi\left(\frac{x}{\sigma}, \frac{t}{\sigma}\right)\big(f_\rho(x) - f_\lambda^\sigma(x)\big)K_u(x)\,d\rho_X(x)\,d\rho_X(t)\right\} = \big(L_K^\sigma\big(f_\rho - f_\lambda^\sigma\big)\big)(u).$$

By (3.4), we see that $\lambda f_\lambda^\sigma = L_K^\sigma(f_\rho - f_\lambda^\sigma) = \mathbb{E}(F(\mathbf{z}, \mathbf{x}))$. Therefore,

$$\left\|f_{\mathbf{z},\lambda}^\sigma - f_\lambda^\sigma\right\|_K \leqslant \frac{1}{\lambda}\left\|F(\mathbf{z}, \mathbf{x}) - \mathbb{E}\big(F(\mathbf{z}, \mathbf{x})\big)\right\|_K.$$

When $k \in \{1, \ldots, m\}$, we see that $F(\mathbf{z}, \mathbf{x}) - \mathbb{E}_{z_k}(F(\mathbf{z}, \mathbf{x}))$ equals

$$\begin{aligned} &\frac{1}{m(m+n)\sigma^d}\left\{\sum_{j\neq k}\Phi\left(\frac{x_k}{\sigma}, \frac{x_j}{\sigma}\right)\big(y_k - f_\lambda^\sigma(x_k)\big)K_u(x_k) - \int_X \Phi\left(\frac{x}{\sigma}, \frac{x_j}{\sigma}\right)\big(f_\rho(x) - f_\lambda^\sigma(x)\big)K_u(x)\,d\rho_X(x)\right\} \\ &\quad + \frac{1}{m(m+n)\sigma^d}\sum_{i\neq k}\big(y_i - f_\lambda^\sigma(x_i)\big)K_u(x_i)\left(\Phi\left(\frac{x_i}{\sigma}, \frac{x_k}{\sigma}\right) - \int_X \Phi\left(\frac{x_i}{\sigma}, \frac{x}{\sigma}\right)d\rho_X(x)\right) \\ &\quad + \frac{1}{m(m+n)\sigma^d}\left\{\Phi\left(\frac{x_k}{\sigma}, \frac{x_k}{\sigma}\right)\big(y_k - f_\lambda^\sigma(x_k)\big)K_u(x_k) - \int_X \Phi\left(\frac{x}{\sigma}, \frac{x}{\sigma}\right)\big(f_\rho(x) - f_\lambda^\sigma(x)\big)K_u(x)\,d\rho_X(x)\right\}. \end{aligned}$$

The reproducing property of RKHS implies that

$$\|f\|_{L_{\rho_X}^2} \leqslant \|f\|_\infty \leqslant \kappa\|f\|_K, \quad \text{for } \forall f \in \mathcal{H}_K. \tag{3.7}$$

Hence, when $k \in \{1, \dots, m\}$, we have

$$\left\| F(\mathbf{z}, \mathbf{x}) - \mathbb{E}_{z_k}\big(F(\mathbf{z}, \mathbf{x})\big) \right\|_K \leqslant \widetilde{M} := \frac{6c_q\kappa(M + \|f_\lambda^\sigma\|_K)}{m\sigma^d}. \tag{3.8}$$

Clearly, the same conclusion also holds for the case $k \in \{m+1, \dots, m+n\}$.

Next we need to estimate the variance $\widetilde{\sigma}^2$, $\|F(\mathbf{z}, \mathbf{x}) - \mathbb{E}_{z_k}(F(\mathbf{z}, \mathbf{x}))\|_K$ can be bounded by

$$\frac{3c_q\kappa(M + \|f_\lambda^\sigma\|_K)}{m(m+n)\sigma^d} \left\{ \sum_{j=1}^{m+n} \left| \Phi\left(\frac{x_k}{\sigma}, \frac{x_j}{\sigma}\right) \right| + \sum_{i \neq k} \left| \Phi\left(\frac{x_i}{\sigma}, \frac{x_k}{\sigma}\right) \right| \right\}.$$

Note that from Definition 1 of Parzen Windows, it follows that $(\mathbb{E}_k(\|F(\mathbf{z}, \mathbf{x}) - \mathbb{E}_{z_k}(F(\mathbf{z}, \mathbf{x}))\|_K^2))^{1/2}$ is bounded by

$$\frac{6c_q\kappa(M + \|f_\lambda^\sigma\|_K)}{m(m+n)\sigma^d} \sum_{j=1}^{m+n} \left(\int_X \frac{c_q^2}{(1 + |(x_j - x)/\sigma|)^{2q}} p(x)\, dx \right)^{1/2} \tag{3.9}$$

$$\leqslant \frac{6c_q^2\kappa c_p(M + \|f_\lambda^\sigma\|_K)}{m\sigma^{d/2}}. \tag{3.10}$$

It follows that for $\sigma \leqslant 1$,

$$\widetilde{\sigma}^2 \leqslant \frac{36c_q^4\kappa^2c_p^2(M + \|f_\lambda^\sigma\|_K)^2}{m\sigma^d}.$$

Thus Proposition 3 follows from Lemma 1. □

4. Approximation error estimate

To estimate the approximation error, we need to consider the convergence of L_K^σ as $\sigma \to 0$.

Proposition 2. *Under Assumption* 1, *then* $V_p \leqslant c_p$ *and for any* $0 < \sigma \leqslant 1$ *we have*

$$\left\| L_K^\sigma - V_pL_K \right\|_{L^2_{\rho_X} \to L^2_{\rho_X}} \leqslant \sigma^\tau\kappa^2c_qc_p(M_{q+2} + M_{q+\tau} + M_q).$$

Here M_{q+2}, $M_{q+\tau}$ *and* M_q *are constants independent of* σ.

Proof. Let $f \in L^2_{\rho_X}$, by Definition 1 we see that $\|L_K^\sigma f - V_pL_Kf\|_{L^2_{\rho_X}}$ is bounded by

$$\frac{\kappa^2}{\sigma^d}\left\{ \int_X\int_X \left| \Phi\left(\frac{x}{\sigma}, \frac{t}{\sigma}\right) \right| \left|f(x)\right| \left|p(x) - p(t)\right| d\rho_X(x)\,dt + \int_{R^d\setminus X}\int_X \left| \Phi\left(\frac{x}{\sigma}, \frac{t}{\sigma}\right) \right| \left|f(x)\right| \left|(p(x)\right| d\rho_X(x)\,dt \right\},$$

and it follows that

$$\frac{\kappa^2}{\sigma^d}\int_X\int_X \left| \Phi\left(\frac{x}{\sigma}, \frac{t}{\sigma}\right) \right| \left|f(x)\right| \left|p(x) - p(t)\right| d\rho_X(x)\,dt \leqslant \frac{\kappa^2}{\sigma^d}\int_X\int_X \frac{c_pc_q}{(1 + |x-t|/\sigma)^q}|x-t|^\tau \left|f(x)\right| dt\,d\rho_X(x)$$
$$\leqslant \sigma^\tau\kappa^2c_pc_qM_{q+\tau}\|f\|_{L^2_{\rho_X}},$$

where $M_{q+\tau}$ is some constant independent of σ and the last inequality is obtained by the Cauchy–Schwartz inequality.

Separate the domain X into $X_\sigma := \{x \in X\colon \inf_{u \in R^d\setminus X}|u - x| \leqslant \sqrt{\sigma}\}$ and its complement $X \setminus X_\sigma$.

When $x \in X \setminus X_\sigma$, for any $t \in R^d \setminus X$, $|t - x| \geqslant \sqrt{\sigma}$ is satisfied, and thereby $1 \leqslant \frac{|t-x|^2}{\sigma}$. Hence

$$\frac{\kappa^2}{\sigma^d}\int_{X\setminus X_\sigma}\int_{R^d\setminus X} \left| \Phi\left(\frac{x}{\sigma}, \frac{t}{\sigma}\right) \right| \left|f(x)\right| \left|p(x)\right| d\rho_X(x)\,dt \leqslant \sigma\kappa^2c_qc_pM_{q+2}\|f\|_{L^2_{\rho_X}},$$

where M_{q+2} is also a constant similar with $M_{q+\tau}$.

For the subset X_σ, we use the Cauchy–Schwartz inequality and obtain

$$\frac{\kappa^2}{\sigma^d}\int_{X_\sigma}\int_{R^d\setminus X} \left| \Phi\left(\frac{x}{\sigma}, \frac{t}{\sigma}\right) \right| \left|f(x)\right| \left|p(x)\right| d\rho_X(x)\,dt \leqslant \sqrt{\rho_X(X_\sigma)}\kappa^2c_qc_pM_q\|f\|_{L^2_{\rho_X}}.$$

By (1.3), $\rho_X(X_\sigma) \leqslant c_p^2\sigma^{2\tau}$. Thus, for any $0<\sigma\leqslant 1$, we have

$$\left\|L_K^\sigma f - V_pL_Kf\right\|_{L^2_{\rho_X}} \leqslant \sigma^\tau\kappa^2c_qc_p(M_{q+2}+M_{q+\tau}+M_q)\|f\|_{L^2_{\rho_X}}.$$

This proves our desired result. □

Define the regularizing function independent on σ

$$f_\lambda = \left(\frac{\lambda}{V_p}I + L_K\right)^{-1} L_Kf_\rho.$$

Proposition 3. *Under the assumptions* (1.3) *and* (1.4)*. When* $L_K^{-r}f_\rho \in L^2_{\rho_X}$ *with* $0<r\leqslant 1$, *we have*

$$\left\|f_\lambda^\sigma - f_\lambda\right\|_{L^2_{\rho_X}} \leqslant C_0\sigma^\tau\lambda^{r-1},$$

where $C_0 = \kappa^2c_qc_pV_p^{r-1}(M_{q+2}+M_{q+\tau}+M_q)\|L_K^{-r}f_\rho\|_{L^2_{\rho_X}}$.

Proof. Observe that

$$f_\lambda^\sigma - f_\lambda = \lambda\big(\lambda I + L_K^\sigma\big)^{-1}\big(L_K^\sigma - V_pL_K\big)(\lambda I + V_pL_K)^{-1}f_\rho.$$

It follows that

$$\left\|f_\lambda^\sigma - f_\lambda\right\|_{L^2_\nu} \leqslant \left\|L_K^\sigma - V_pL_K\right\|\left\|(\lambda I + V_pL_K)^{-1}L_K^r\right\|\left\|L_K^{-r}f_\rho\right\|_{L^2_{\rho_X}} \leqslant \lambda^{r-1}\left\|L_K^\sigma - V_pL_K\right\|\left\|L_K^{-r}f_\rho\right\|_{L^2_{\rho_X}}.$$

The desired result follows from Proposition 3. □

To get the total error $\|f_{\mathbf{z},\lambda}^\sigma - f_\rho\|$ we need bounds for the approximation error $\|f_\lambda - f_\rho\|$. The following result is well known, proved by Smale and Zhou in [16].

Lemma 2. *Define* f_λ *as above. If* $L_K^{-r}f_\rho \in L^2_{\rho_X}(X)$ *with* $0<r\leqslant 1$, *then*

$$\|f_\lambda - f_\rho\|^2_{L^2_{\rho_X}} + \lambda\|f_\lambda\|_K^2 \leqslant \lambda^{2r}V_p^{-2r}\left\|L_K^{-r}f_\rho\right\|^2_{L^2_{\rho_X}}, \quad \textit{if}\ 0<r<\frac{1}{2} \tag{4.1}$$

and

$$\|f_\lambda - f_\rho\|_K \leqslant \lambda^{r-\frac{1}{2}}V_p^{\frac{1}{2}-r}\left\|L_K^{-r}f_\rho\right\|_{L^2_{\rho_X}}, \quad \textit{if}\ \frac{1}{2}\leqslant r\leqslant 1. \tag{4.2}$$

Moreover, for $0<r\leqslant 1$, *there holds*

$$|f_\lambda - f_\rho\|_{L^2_{\rho_X}} \leqslant \lambda^rV_p^{-r}\left\|L_K^{-r}f_\rho\right\|_{L^2_{\rho_X}}. \tag{4.3}$$

Following Proposition 3 and (4.3), the approximation error can be stated as follows.

Proposition 4. *Under the assumptions* (1.3) *and* (1.4)*. When* $L_K^{-r}f_\rho \in L^2_{\rho_X}$ *with* $0<r\leqslant 1$, *we have*

$$\left\|f_\lambda^\sigma - f_\rho\right\|_{L^2_{\rho_X}} \leqslant \big(\sigma^\tau\lambda^{r-1} + \lambda^r\big)\big(C_0 + V_p^{-r}\left\|L_K^{-r}f_\rho\right\|_{L^2_{\rho_X}}\big), \tag{4.4}$$

where C_0 *is defined as in Proposition* 3.

5. Learning rate and conclusion

We are now in a position to derive the learning rate of Semi-Supervised Learning algorithm (1.2).

Proof of Theorem 1. Following Propositions 3 and 4, with confidence to $1-\delta$,

$$\begin{aligned}\left\|f_{\mathbf{z},\lambda}^\sigma - f_\rho\right\|_{L^2_{\rho_X}} &\leqslant \left\|f_{\mathbf{z},\lambda}^\sigma - f_\lambda^\sigma\right\|_{L^2_{\rho_X}} + \left\|f_\lambda^\sigma - f_\rho\right\|_{L^2_{\rho_X}}\\ &\leqslant \log\left(\frac{2}{\delta}\right)\frac{12c_q\kappa^2(M + \|f_\lambda^\sigma\|_K)}{\sqrt{m}\lambda\sigma^{d/2}}\left(\frac{1}{\sqrt{m}\sigma^{d/2}} + c_qc_p\right) + \big(\sigma^\tau\lambda^{r-1} + \lambda^r\big)\big(C_0 + V_p^{-r}\left\|L_K^{-r}f_\rho\right\|_{L^2_{\rho_X}}\big)\\ &\leqslant \widetilde{C}\log\left(\frac{2}{\delta}\right)\left(\frac{1}{m}\right)^{\frac{\tau r}{2\tau(r+1)+d}},\end{aligned}$$

provided that $\lambda = \sigma^\tau$ and $\lambda = (\frac{1}{m})^{\frac{\tau}{2\tau(r+1)+d}}$, where $\widetilde{C}$ is a constant independent on m or δ.

We introduce a Semi-Supervised Learning algorithm with density estimators, generated by Parzen Windows functions. The main difference from Weighted Supervised Learning is that large amounts of unlabeled data are employed in density estimations. Thus estimation parameter may be less sensitive than the case where only a few observations are used, such as Weighted Least Squared Regression. An error analysis is given for the convergence of the learner to the regression function. The technique we use here is capacity independent one. Considering the complexity of hypothesis space such as Rademacher complexity [18] and various covering numbers [20], we are sure that the convergence rate can be improved greatly if the hypothesis space is good enough. □

References

[1] M.F. Balcan, A. Blum, A discriminative model for semi-supervised learning, J. ACM (2009).

[2] M. Belkin, I. Matveeva, P. Niyogi, Regularization and semi-supervised learning on large graphs, in: COLT, 2004.

[3] O. Chapelle, A. Zien, B. Scholkopf, Semi-Supervised Learning, MIT Press, 2006.

[4] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. (N.S.) 39 (2001) 1–49.

[5] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, Cambridge, 2007.

[6] L. Devroye, G. Lugosi, Combinatorial Methods in Density Estimation, Springer, Heidelberg, 2000.

[7] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math. 13 (2000) 1–50.

[8] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic, London, 1990.

[9] Rie Kubota Ando, Tong Zhang, Two-view feature generation model for semi-supervised, in: Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007.

[10] X. Lafferty, J. Zhu, Y. Liu, Kernel conditional random fields: Representation and clique selection, in: The 21st International Conference on Machine Learning (ICML), 2004.

[11] S. Mukherjee, D.X. Zhou, Learning coordinate covariances via gradients, J. Mach. Learn. Res. 7 (2006) 519–549.

[12] K. Nigam, T. Mitchell, A. McCallum, S. Thrun, Text classification from labeled and unlabeled documents using EM, in: Machine Learning, Kluwer Academic Publishers, Boston, 2000, pp. 1–34.

[13] E. Parzen, On the estimation of a probability density function and the mode, Ann. Math. Stat. 33 (1962) 1049–1051.

[14] Z.W. Pan, D.H. Xiao, Q.W. Xiao, D.X. Zhou, Parzen windows for multi-class classification, J. Complexity 24 (2008) 606–618.

[15] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, Ann. Probab. 22 (1994) 1679–1706.

[16] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.

[17] B. Taskar, C. Guestrin, D. Koller, Max-margin Markov networks, in: NIPS'03, 2003.

[18] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer-Verlag, New York, 1996.

[19] M.P. Wand, M.C. Jones, Kernel Smoothing, Monogr. Statist. Appl. Probab., vol. 60, Chapman Hall, London, 1995.

[20] D.X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1743–1752.

[21] X.J. Zhou, D.X. Zhou, High order Parzen windows and randomized sampling, Adv. Comput. Math. 31 (2009) 349–368.

[22] X.J. Zhu, Semi-supervised learning literature survey, Tech. Report, 1530.

[23] X.J. Zhu, Semi-supervised learning with graph, Doctoral Thesis, Carnegie Mellon University, 2005.