



高斯核正则化学习算法的泛化误差 *

^{1,2} 张永全 ¹ 李有梅

(¹ 中国计量学院数学与信息科学系 杭州 310018; ² 西安交通大学信息与系统科学研究所 西安 710049)

摘要: 对广义凸损失函数和变高斯核情形下正则化学习算法的泛化性能展开研究. 其目标是给出学习算法泛化误差的一个较为满意上界. 泛化误差可以利用正则误差和样本误差来测定. 基于高斯核的特性, 通过构建一个径向基函数 (简记为 RBF) 神经网络, 给出了正则误差的上界估计, 通过投影算子和再生高斯核希尔伯特空间的覆盖数给出样本误差的上界估计. 所获结果表明, 通过适当选取参数 σ 和 λ , 可以提高学习算法的泛化性能.

关键词: 学习理论; RBF 神经网络; 高斯核; 泛化误差.

MR(2000) 主题分类: 41A25; 41A63; 42B08 **中图分类号:** O174.41 **文献标识码:** A

文章编号: 1003-3998(2014)05-1049-12

1 引言

学习算法的泛化性能分析是当前学习理论研究的热点之一. Vapnik and Chervonenkis^[1] 首先从统计学的观点研究了分类学习算法的泛化性能. 随后, 多种不同的理论工具被用于学习算法的泛化性能研究中, 尤其是分类 (二值函数学习) 和回归 (实值函数学习) 学习问题. 本文根据 RBF 神经网络的性质和再生高斯核希尔伯特空间的覆盖数理论, 对广义凸损失函数和变高斯核情形下的正则化学习算法进行理论分析, 得到正则化学习算法泛化误差的上界估计. 目前已有不少学者研究了学习算法的泛化性能, 如文献 [2-4]. 然而, 由于未知范数的等价关系, 泛化误差研究成为一个公认的难题, 自然会问, 如何估计泛化误差的上界? 本文通过估计变高斯核再生核希尔伯特空间的范数, 及利用高斯核函数的特殊性质, 构造目标函数一个好的逼近, 得到泛化误差的上界估计. 具体来说, 利用高斯函数构建可逼近 $[-1, 1]^d$ 上连续函数的一个 RBF 神经网络, 并将逼近结果应用于泛化误差的上界估计.

众所周知, RBF 神经网络是一个万能逼近器, 即, 紧致集上的任意连续函数都可以用 RBF 神经网络来逼近 (见文献 [5-6]). 然而, 他们仅考虑了 RBF 的收敛性质. 对多变量连续函数有些学者利用相关算子研究了收敛速率, 然而在参数选择和算法实现上都存在困难. 本文构建了一个具体的高斯 RBF 神经网络, 其参数可以计算得到.

令 $X = [-1, 1]^d$ 和 $Y = \{y \in \mathbb{R}, |y| \leq M\} \subset \mathbb{R}$, 其中 $M > 0$. 设 ρ 是 $Z = X \times Y$ 上的一个未知概率分布. Z^n 表示由 n 个依据分布 ρ 独立抽取随机样本点的集合. 用 $\rho_X, \rho(y|x)$ 分别表示 ρ 的边缘概率测度和条件概率测度. 设核函数 $K: X \times X \rightarrow \mathbb{R}$ 是连续、对称、半正定

收稿日期: 2012-07-28; 修订日期: 2013-12-27

E-mail: zymath@163.com

* 基金项目: 国家自然科学基金 (11301494) 和浙江省自然科学基金 (Q12A01026) 资助

的, 即, 对于 $\{x_1, x_2, \dots, x_l\} \subset X$, 矩阵 $(K(x_i, x_j))_{i,j=1}^l$ 是半正定的. 称此核为 Mercer 核. 再生核希尔伯特空间 (RKHS) $\mathcal{H}_K$ (见文献 [7-12]), 定义为函数集 $\{K_x = K(x, \cdot) : x \in X\}$ 的线性组合张成的闭包. 用 $\langle \cdot, \cdot \rangle_K$ 表示内积, $\langle K_x, K_y \rangle_K = K(x, y)$, 称

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in X, f \in \mathcal{H}_K \quad (1.1)$$

为核 K 的再生性.

令 $\mathcal{C}(X)$ 表示 X 上的连续函数集合. 由 (1.1) 式知

$$\|f\|_\infty \leq \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K. \quad (1.2)$$

其中 $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$.

Mercer 核已经被应用于许多学习算法中, 而高斯核 $K_\sigma(x, y) = \exp\{-\|x - y\|^2/(2\sigma^2)\}$ 作为可变核的一个特例. Xiang 和 Zhou^[13] 在 Sobolev 光滑条件下, 利用傅里叶分析研究了变高斯核正则化分类算法的收敛速率. 文献 [14] 分析了可变核学习算法的收敛性能. 尤其在回归函数满足 Lipschitz 条件时, Ye 和 Zhou 在文献 [15] 中研究了黎曼子流形 $\mathbb{R}^d$ 上变高斯核学习算法的泛化性能. Maiorov^[16] 利用假设空间的熵和网络的逼近性能研究神经网络学习算法的收敛性. 本文在目标函数满足 Lipschitz 条件下研究高斯 RBF 神经网络正则化学习算法的泛化性能.

记

$$\mathcal{H}_\sigma = \overline{\text{span}\{K_\sigma(x, \cdot) : x \in [-1, 1]^d\}}.$$

定义 1.1 设 $\ell : \mathbb{R} \times X \times Y \rightarrow [0, \infty)$ 为一个广义凸损失函数. 如果

$$L = \sup_{f, f' \in B_M, f \neq f'} \sup_{x \in X, y \in Y} \frac{|\ell(f, x, y) - \ell(f', x, y)|}{\|f - f'\|_\infty} < \infty,$$

则称 ℓ 是局部 Lipschitz 连续的函数, 其中, $B_M = \{f : \sup_{x \in X} |f(x)| \leq (4d+1)M + 5dC + 13\}$, C 为 Lipschitz 常数.

函数 $f : X \rightarrow Y$ 的泛化误差定义为

$$\mathcal{E}(f) = \int_Z \ell(f, x, y) d\rho. \quad (1.3)$$

我们在所有可测函数集合中把使得 (1.3) 式达到最小的函数记为 f_ρ^* , 即

$$f_\rho^* = \arg \min_{f: X \rightarrow \mathbb{R}} \int_{X \times Y} \ell(f, x, y) d\rho.$$

本文假设函数 f_ρ^* 存在且损失函数 ℓ 是关于 $f(x)$ 的单调递增函数.

假定 (Zhou 和 Jetter^[11]) 存在 $\beta \in [0, 1]$ 和常数 $C_1 > 0$ 使得下式成立

$$\mathbf{E} \{(\ell(f, x, y) - \ell(f_\rho^*, x, y))^2\} \leq C_1 (\mathcal{E}(f) - \mathcal{E}(f_\rho^*))^\beta, \quad \forall f \in \mathcal{H}_\sigma.$$

对于 $\beta = 0$ 和 $C_1 = M^2$, 这个假定总是成立的. 本文仅考虑 $\beta = 1$ 的情形.

定义可变核正则化学习算法

$$f_{\mathbf{z}, \sigma} = f_{\mathbf{z}, \sigma, \lambda} = \arg \min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(f, x_i, y_i) + \lambda \|f\|_{K_\sigma}^2 \right\}.$$

函数 f 的经验误差定义为

$$\mathcal{E}_z(f) = \frac{1}{n} \sum_{i=1}^n \ell(f, x_i, y_i).$$

因此

$$f_{z,\sigma} = f_{z,\sigma,\lambda} = \arg \min_{f \in \mathcal{H}_\sigma} \{ \mathcal{E}_z(f) + \lambda \|f\|_{K_\sigma}^2 \}, \quad (1.4)$$

其中 $\lambda \geq 0$ 为正则化参数. 该参数通常依赖于样本容量 n , 即, $\lambda = \lambda(n)$, $\lim_{n \rightarrow 0} \lambda(n) = 0$. 对于 $x \in X$, 我们总假设 $|f_\rho^*(x)| \leq M$.

定义 1.2 (Wu, Ying 和 Zhou^[2]) 对于可测函数空间 $f: X \rightarrow \mathbb{R}$, 定义其投影算子如下

$$\pi_M(f)(x) = \begin{cases} M, & f(x) > M, \\ f(x), & -M \leq f(x) \leq M, \\ -M, & f(x) < -M. \end{cases}$$

算法 (1.4) 的有效性是通过逼近函数 $\pi_M(f_{z,\sigma})$ 和最优函数 f_ρ^* 之间的均方误差来度量. 为了界定泛化误差, 需要定义残差

$$\int_Z \{ \ell(\pi_M(f_{z,\sigma}), x, y) - \ell(f_\rho^*, x, y) \} d\rho = \mathcal{E}(\pi_M(f_{z,\sigma})) - \mathcal{E}(f_\rho^*).$$

本文目标是理解学习算法 (1.4) 的泛化性能以及对应参数 σ 的选取. 文献 [17] 通过选取合适的参数 σ 研究了支持向量机分类器的泛化性能.

为了获得泛化误差的上界估计, 函数 f_ρ^* 需满足一些光滑性条件.

对于 $C > 0, x, y \in [-1, 1]^d$, 若函数 $f: X \rightarrow \mathbb{R}$ 满足

$$|f(x) - f(y)| \leq C \|x - y\|.$$

则称 $f \in \text{Lip}_C 1$. 对 $t \geq 0$, 定义函数 f 的连续模

$$\omega(f, t) = \sup_{x, y \in [-1, 1]^d, \|x - y\| \leq t} |f(x) - f(y)|.$$

由 $f \in \text{Lip}_C 1$, 推知

$$\omega(f, t) \leq Ct.$$

众所周知, 高斯 RBF 神经网络在许多领域扮演了重要的角色, 其定义如下

$$G_m(x) = \sum_{J \in M_m} \beta_J \exp \left\{ -\frac{m \|x - \alpha_J\|^2}{4} \right\},$$

其中, $M_m = \{J = (j_1, j_2, \dots, j_d) : j_k = 0, 1, 2, \dots, m-1, k = 1, 2, \dots, d\}$, $\beta_J \in \mathbb{R}, \alpha_J \in [-1, 1]^d$. 明显地, 当 $\sigma = \frac{2}{\sqrt{m}}$ 时, $G_m(\cdot) \in \mathcal{H}_\sigma$.

本文结构组织如下. 第二节将介绍正则误差, 并利用高斯 RBF 神经网络给出其上界估计. 在第三节给出样本误差的上界估计, 并综合正则误差的上界估计, 给出泛化误差的上界估计. 最后, 总结文中所得的结论.

2 正则误差和逼近

对于 $\sigma^2 = \frac{4}{m}$, 在 $\mathcal{H}_\sigma$ 上构建一个 RBF 神经网络 $G_m(f_\rho^*)(x)$

$$G_m(f_\rho^*)(x) = \sum_{J \in M_m} \left(3\sqrt{\frac{1}{4m\pi}} \right)^d f_\rho^*(a_J) \exp \left\{ -\frac{m\|x - a_J\|_2^2}{4} \right\}, \quad (2.1)$$

去估计泛化误差. 其中 $a_J = \frac{2}{m}J + (-1, -1, \dots, -1)$.

引理 2.1 将残差 $\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*)$ 分解为

$$\begin{aligned} & \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) \\ & \leq \{ \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_\mathbf{z}(\pi_M(f_{\mathbf{z},\sigma})) + \mathcal{E}_\mathbf{z}(G_m(f_\rho^*)) - \mathcal{E}(G_m(f_\rho^*)) \} + \mathcal{D}(\lambda), \end{aligned} \quad (2.2)$$

其中 $\mathcal{D}(\lambda) = \mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) + \lambda \|G_m(f_\rho^*)\|_{K_\sigma}^2$.

证 由 (1.4) 式和定义 1.2 知

$$\begin{aligned} & \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) \leq \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) + \lambda \|f_{\mathbf{z},\sigma}\|_{K_\sigma}^2 \\ & = \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_\mathbf{z}(\pi_M(f_{\mathbf{z},\sigma})) + \mathcal{E}_\mathbf{z}(\pi_M(f_{\mathbf{z},\sigma})) + \lambda \|f_{\mathbf{z},\sigma}\|_{K_\sigma}^2 - \mathcal{E}_\mathbf{z}(G_m(f_\rho^*)) \\ & \quad - \lambda \|G_m(f_\rho^*)\|_{K_\sigma}^2 + \mathcal{E}_\mathbf{z}(G_m(f_\rho^*)) - \mathcal{E}(G_m(f_\rho^*)) + \mathcal{D}(\lambda) \\ & \leq \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_\mathbf{z}(\pi_M(f_{\mathbf{z},\sigma})) + \mathcal{E}_\mathbf{z}(G_m(f_\rho^*)) - \mathcal{E}(G_m(f_\rho^*)) + \mathcal{D}(\lambda). \end{aligned}$$

证毕. |

称 (2.2) 式的第一项为样本误差, 第二项为正则误差. 为估计正则误差, 需要分别估计 $\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*)$ 和 $\lambda \|G_m(f_\rho^*)\|_{K_\sigma}^2$.

定理 2.1 设 $f_\rho^* \in \text{Lip}_C 1$ 和 $|f_\rho^*(x)| \leq M$, 则对任意的正整数 m 和所有的 $x \in [-1, 1]^d$, 有

$$|f_\rho^*(x) - G_m(f_\rho^*, x)| \leq \frac{5dC + 4dM + 13}{\sqrt{m}}.$$

为了证明定理 2.1, 需要引入引理 2.2 和引理 2.3.

引理 2.2 (Xie 和 Cao^[18]) 对 $f \in C_{[-1,1]^d}$, 则存在 $F \in C(\mathbb{R}^d)$ 满足

$$F(x) = f(x), \quad x \in [-1, 1]^d,$$

而且对于 $x, y \in [-1, 1]^d$, $\|x - y\|_2 < \delta \leq 1$, 有

$$\|F\|_\infty = \max_{x \in \mathbb{R}^d} |F(x)| \leq \|f\|_\infty = \max_{x \in [-1, 1]^d} |f(x)|, \quad \|F(x) - F(y)\| \leq \omega(f, \delta).$$

引理 2.2 表明, 存在 $F_\rho \in C(\mathbb{R}^d)$ 满足

$$F_\rho(x) = f_\rho^*(x), \quad x \in [-1, 1]^d,$$

且对任意 $0 < \delta \leq 1$, 满足

$$\|F_\rho\|_\infty \leq \|f_\rho^*\|_\infty, \quad \omega(F_\rho, \delta) \leq \omega(f_\rho^*, \delta).$$

引理 2.3 (Xie 和 Cao^[18]) 对于 $F_\rho \in C(\mathbb{R}^d)$ 和 $x \in \mathbb{R}^d$, 则有

$$|H_m(F_\rho, x) - F_\rho(x)| \leq \left(1 + \frac{2}{\sqrt{\pi}} \right) d\omega \left(f_\rho^*, \frac{1}{\sqrt{m}} \right).$$

其中算子 $H_m(F_\rho, \cdot)$ 定义如下

$$H_m(F_\rho, x) = \left(\frac{m}{4\pi}\right)^{\frac{d}{2}} \int_{\mathbb{R}^d} F_\rho(x-y) \exp\left\{-\frac{m\|y\|^2}{4}\right\} dy.$$

引理 2.4 (Xie 和 Cao^[18]) 对于 $F_\rho \in C(\mathbb{R}^d)$, 有

$$|H_m(F_\rho, x) - G_m(F_\rho, x)| \leq 2d\omega\left(f_\rho^*, \frac{2}{m}\right) + \frac{13}{\sqrt{m}} + \frac{4d}{\sqrt{\pi m}} \|F_\rho\|.$$

定理 2.1 的证明 由引理 2.2、引理 2.3、引理 2.4 得

$$\begin{aligned} & |f_\rho^*(x) - G_m(f_\rho^*, x)| \\ & \leq |f_\rho^*(x) - H_m(f_\rho^*, x)| + |H_m(f_\rho^*, x) - G_m(f_\rho^*, x)| \\ & \leq \left(1 + \frac{2}{\sqrt{\pi}}\right) d\omega\left(F_\rho, \frac{1}{\sqrt{m}}\right) + 2d\omega\left(f_\rho^*, \frac{2}{m}\right) + \frac{13}{\sqrt{m}} + \frac{4d}{\sqrt{\pi m}} \|F_\rho\| \\ & \leq 3d\omega\left(f_\rho^*, \frac{1}{\sqrt{m}}\right) + 2d\omega\left(f_\rho^*, \frac{2}{m}\right) + \frac{13}{\sqrt{m}} + \frac{4dM}{\sqrt{\pi m}}. \end{aligned}$$

由于 $f_\rho^* \in \text{Lip}_C 1$, 则有

$$|f_\rho^*(x) - G_m(f_\rho^*, x)| \leq \frac{5dC + 4dM + 13}{\sqrt{m}}.$$

证毕. |

定理 2.2 设 m 为一个正整数, $X = [-1, 1]^d$. $G_m(f_\rho^*)$ 由 (2.1) 式给出. 则对 $f_\rho^* \in C_{[-1, 1]^d}$ 和 $|f_\rho^*(x)| \leq M$, 有

$$\|G_m(f_\rho^*)\|_{K_\sigma} \leq \left(\frac{9m}{4\pi}\right)^d M^2.$$

证 因为 $G_m(f_\rho^*) \in \mathcal{H}_\sigma$, 其中 $\sigma^2 = \frac{4}{m}$, 则有

$$\begin{aligned} & \|G_m(f_\rho^*)\|_{K_\sigma}^2 = \langle G_m(f_\rho^*), G_m(f_\rho^*) \rangle_{K_\sigma} \\ & = \left(3\sqrt{\frac{1}{4m\pi}}\right)^{2d} \left\langle \sum_{J \in M_m} f_\rho^*(a_J) \exp\left\{-\frac{m\|x - a_J\|_2^2}{4}\right\}, \sum_{J \in M_m} f_\rho^*(a_J) \exp\left\{-\frac{m\|x - a_J\|_2^2}{4}\right\} \right\rangle_{K_\sigma} \\ & = \left(\frac{9}{4m\pi}\right)^d \sum_{I, J \in M_m} f_\rho^*(a_J) f_\rho^*(a_I) \exp\left\{-\frac{m\|a_I - a_J\|_2^2}{4}\right\}. \end{aligned}$$

于是可得

$$\begin{aligned} \|G_m(f_\rho^*)\|_{K_\sigma}^2 & \leq \left(\frac{9}{4m\pi}\right)^d M^2 \sum_{I, J \in M_m} \exp\left\{-\frac{m\|a_I - a_J\|_2^2}{4}\right\} \\ & = \left(\frac{9}{4m\pi}\right)^d M^2 \sum_{I, J \in M_m} 1 \leq \left(\frac{9}{4m\pi}\right)^d M^2 \times m^{2d} = \left(\frac{9m}{4\pi}\right)^d M^2. \end{aligned}$$

证毕. |

从定理 2.1 知, 对任意 $x \in X$, 有

$$|G_m(f_\rho^*, x)| \leq |f_\rho^*(x)| + \frac{5dC + 4dM + 13}{\sqrt{m}} \leq M + \frac{5dC + 4dM + 13}{\sqrt{m}}$$

和

$$\|G_m(f_\rho^*, \cdot)\|_\infty \leq M + \frac{5dC + 4dM + 13}{\sqrt{m}} \leq (4d + 1)M + 5dc + 13.$$

因此有 $G_m(f_\rho^*, \cdot) \in B_M$.

结合定义 1.1, 可推知

$$\begin{aligned} \mathcal{E}(G_m(f_\rho^*, \cdot)) - \mathcal{E}(f_\rho^*) &= \int_Z (\ell(G_m(f_\rho^*), x, y) - \ell(f_\rho^*, x, y)) d\rho \\ &\leq \int_Z L \|G_m(f_\rho^*, \cdot) - f_\rho^*\|_\infty \leq \frac{5dC + 4dM + 13}{\sqrt{m}} L. \end{aligned} \quad (2.3)$$

3 样本误差估计

为估计泛化误差 $\mathcal{E}(\pi_M(f_{z, \sigma})) - \mathcal{E}(f_\rho^*)$ 的上界, 我们还需要给出样本误差的上界估计. 根据 (2.2) 式的第一部分知

$$\begin{aligned} &\mathcal{E}(\pi_M(f_{z, \sigma})) - \mathcal{E}_z(\pi_M(f_{z, \sigma})) + \mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}(G_m(f_\rho^*)) \\ &\leq |\mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}_z(f_\rho^*) - \mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*)| \\ &\quad + |\mathcal{E}(\pi_M(f_{z, \sigma})) - \mathcal{E}(f_\rho^*) - \mathcal{E}_z(\pi_M(f_{z, \sigma})) - \mathcal{E}_z(f_\rho^*)|. \end{aligned} \quad (3.1)$$

引理 3.1 (Cucker 和 Zhou^[19]) 设 ρ 是定义在 $Z = X \times Y$ 上的概率测度, $z_1 = (x_1, y_1), \dots, z_n = (x_n, y_n)$ 为根据 ρ 所取的样本集. 对于函数 $g: Z \rightarrow \mathbb{R}$, 设 $S = \sum_{i=1}^n g(z_i)$, $b = \|g\|_\infty$, $\sigma^2 = n\mathbf{E}g^2$, 成立

$$\text{Prob}_{z \in Z^n} \{|S - \mathbf{E}S| \geq t\} \leq 2 \exp \left\{ -\frac{t^2}{2(\sigma^2 + \frac{bt}{3})} \right\}.$$

其中 $\mathbf{E}\xi$ 表示随机变量 ξ 的数学期望.

由引理 3.1 可推出如下定理.

定理 3.1 若损失函数 ℓ 满足假定, 则对 $0 < \delta \leq 1$, 不等式

$$|\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}_z(f_\rho^*))| \leq \frac{6C_1 + 2M}{3n} \log \frac{4}{\delta} + \frac{1}{2} D(\lambda)$$

至少以 $1 - \frac{\delta}{2}$ 的概率成立.

证 令 $g(z) = \frac{1}{n} \{\ell(G_m(f_\rho^*), x, y) - \ell(f_\rho^*, x, y)\}$. 对任意 $z \in Z$, 有

$$|g(z)| = \frac{1}{n} |\ell(G_m(f_\rho^*), x, y) - \ell(f_\rho^*, x, y)| \leq \frac{M}{n}.$$

由假定知

$$\begin{aligned} \mathbf{E}g^2 &= \frac{1}{n^2} \mathbf{E}\{(\ell(G_m(f_\rho^*), \cdot, \cdot) - \ell(f_\rho^*, \cdot, \cdot))^2\} \leq \frac{C_1}{n^2} \mathbf{E}\{\ell(G_m(f_\rho^*), \cdot, \cdot) - \ell(f_\rho^*, \cdot, \cdot)\} \\ &= \frac{C_1}{n} \mathbf{E}\left\{\frac{1}{n} \{\ell(G_m(f_\rho^*), \cdot, \cdot) - \ell(f_\rho^*, \cdot, \cdot)\}\right\} = \frac{C_1}{n} \mathbf{E}g. \end{aligned}$$

令 $t = \sqrt{\varepsilon(\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) + \varepsilon)}$. 由引理 3.1 知, 对于任何 $\varepsilon > 0$, 不等式

$$\frac{|\mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}_z(f_\rho^*) - \mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*)|}{\sqrt{(\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) + \varepsilon)}} \leq \sqrt{\varepsilon}$$

至少以

$$\begin{aligned} & 1 - 2 \exp \left\{ - \frac{\varepsilon(\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) + \varepsilon)}{2 \left(\frac{C_1}{n} (\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*)) + \frac{M \sqrt{\varepsilon(\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) + \varepsilon)}}{3n} \right)} \right\} \\ & \geq 1 - 2 \exp \left\{ - \frac{3n\varepsilon}{6C_1 + 2M} \right\} \end{aligned}$$

的概率成立. 其中 $b = \frac{M}{n}$ 和 $\sigma^2 = n\mathbf{E}g^2 \leq C_1\mathbf{E}g$.

由不等式

$$ab \leq \frac{1}{2}(a^2 + b^2) \quad \forall a, b \in \mathbb{R},$$

推知

$$\begin{aligned} |\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}_z(f_\rho^*))| & \leq \frac{1}{2}\varepsilon + \frac{1}{2}(\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) + \varepsilon) \\ & = \varepsilon + \frac{1}{2}D(\lambda). \end{aligned}$$

令 $\frac{\delta}{2} = 2 \exp \left\{ - \frac{3n\varepsilon}{6C_1 + 2M} \right\}$, 则 $\varepsilon = \frac{(6C_1 + 2M) \log \frac{4}{\delta}}{3n}$. 因此

$$|\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}_z(f_\rho^*))| \leq \frac{6C_1 + 2M}{3n} \log \frac{4}{\delta} + \frac{1}{2}D(\lambda)$$

至少以 $1 - \frac{\delta}{2}$ 的概率成立. |

由于函数 $\xi = \ell(\pi_M(f_{\mathbf{z}, \sigma}), x, y) - \ell(f_\rho^*, x, y)$ 是随着样本 $\mathbf{z}$ 的变化而变化的, 所以不能将其看作是一个固定的函数. 为此, 如何估计 (3.1) 式中 $|\mathcal{E}_z(G_m(f_\rho^*)) - \mathcal{E}_z(f_\rho^*) - \mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*)|$ 的上界是学习算法中的难点之一. 在估计中, 我们将要用到覆盖数理论.

定义 3.1 设 $\mathcal{F}$ 为度量空间的一个子集, $\varepsilon > 0$. 我们将半径为 ε 且能够覆盖 $\mathcal{F}$ 圆盘的最小个数定义为 $\mathcal{F}$ 的覆盖数.

关于覆盖数的一些结论可参见文献 [20–24]. 为估计样本误差, 我们还需如下引理.

引理 3.2 (Zhou^[22]) 存在一个仅依赖于 X 和 d 的常数 C_0 , 使得

$$\log \mathcal{N}(B_1, \eta) \leq C_0 \left(\left(\log \frac{1}{\eta} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right), \quad \forall 0 < \eta < 1, 0 < \sigma \leq 1,$$

其中, $B_1 = \{f \in \mathcal{H}_\sigma : \|f\|_{K_\sigma} \leq 1\}$.

引理 3.3 对所有 $\lambda > 0$ 和 $\mathbf{z} \in Z^n$, 下式成立

$$\|f_{\mathbf{z}, \sigma}\|_{K_\sigma} \leq \sqrt{\frac{M}{\lambda}}.$$

证 由 $f_{\mathbf{z}, \sigma}$ 的定义知

$$\lambda \|f_{\mathbf{z}, \sigma}\|_{K_\sigma}^2 \leq \mathcal{E}_z(f_{\mathbf{z}, \sigma}) + \lambda \|f_{\mathbf{z}, \sigma}\|_{K_\sigma}^2 \leq \mathcal{E}_z(0) \leq M.$$

则有

$$\|f_{\mathbf{z},\sigma}\|_{K_\sigma} \leq \sqrt{\frac{M}{\lambda}}.$$

证毕.

定理 3.2 若损失函数 ℓ 满足假定. 则对任何 $\varepsilon > 0$, 有

$$\begin{aligned} & \text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \geq \varepsilon \} \\ & \leq 2 \exp \left\{ C_0 \left(\left(\log \frac{4L\sqrt{M}}{\varepsilon\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right) - \frac{3n(3\varepsilon - D(\lambda))}{6(M + C_1)} \right\}. \end{aligned}$$

证 设 $R = \sqrt{\frac{M}{\lambda}}$, 则 $f_{\mathbf{z},\sigma} \in B_R$. 由 $f_{\mathbf{z},\sigma}$ 的定义知

$$\begin{aligned} & |\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},\sigma})) + \mathcal{E}_{\mathbf{z}}(f_\rho^*)| \\ & \leq \sup_{f \in B_R} |\mathcal{E}(\pi_M(f)) - \mathcal{E}(f_\rho^*) - \mathcal{E}_{\mathbf{z}}(\pi_M(f)) + \mathcal{E}_{\mathbf{z}}(f_\rho^*)|. \end{aligned}$$

由于

$$\begin{aligned} |\ell(\pi_M(f), x, y) - \ell(\pi_M(g), x, y)| & \leq L|\pi_M(f)(x) - \pi_M(g)(x)| \\ & \leq L|f(x) - g(x)|, \quad f, g \in B_R, \end{aligned}$$

因此

$$\begin{aligned} & |\mathcal{E}(\pi_M(f)) - \mathcal{E}(\pi_M(g)) - \mathcal{E}_{\mathbf{z}}(\pi_M(f)) + \mathcal{E}_{\mathbf{z}}(\pi_M(g))| \\ & \leq 2L\|f - g\|_\infty \leq 2L\|f - g\|_K, \quad f, g \in B_R. \end{aligned}$$

用 $U = \{f_1, f_2, \dots, f_l\} \subset B_R$ 表示 B_R 的大小为 $l = \mathcal{N}(B_R, \gamma)$ 一个 γ -网. 则有

$$\begin{aligned} & \sup_{f \in B_R} |\mathcal{E}(\pi_M(f)) - \mathcal{E}_{\mathbf{z}}(\pi_M(f)) - \mathcal{E}(f_\rho^*) + \mathcal{E}_{\mathbf{z}}(f_\rho^*)| \\ & \leq \sup_{f \in U} |\mathcal{E}(\pi_M(f)) - \mathcal{E}_{\mathbf{z}}(\pi_M(f)) - \mathcal{E}(f_\rho^*) + \mathcal{E}_{\mathbf{z}}(f_\rho^*)| + 2L\gamma. \end{aligned}$$

对任意的 $f_i \in U$, 类似于定理 3.1 证明方法, 可得

$$\text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(f_i) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(f_i) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \geq \varepsilon \} \leq 2 \exp \left\{ -\frac{3n(\varepsilon - \frac{1}{2}D(\lambda))}{2(M + 3C_1)} \right\}.$$

于是有

$$\begin{aligned} & \text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(f_{\mathbf{z},\sigma}) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\sigma}) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \geq \varepsilon \} \\ & \leq \text{Prob}_{\mathbf{z} \in Z} \left\{ \sup_{f \in B_R} |\mathcal{E}(f) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \geq \varepsilon \right\} \\ & \leq \text{Prob}_{\mathbf{z} \in Z} \left\{ \sup_{f \in U} |\mathcal{E}(f) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \geq \varepsilon - 2L\gamma \right\} \\ & \leq \mathcal{N}(B_R, \gamma) \sup_{f \in U} \text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(f) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \geq \varepsilon - 2L\gamma \} \\ & \leq 2\mathcal{N}(B_R, \gamma) \exp \left\{ -\frac{3n(\varepsilon + 2L\gamma - \frac{1}{2}D(\lambda))}{2(M + 3C_1)} \right\}. \end{aligned}$$

当 $\gamma = \frac{\varepsilon}{4L}$, 则有

$$\begin{aligned} & \text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(f_{\mathbf{z}, \sigma}) - \mathcal{E}(f_{\rho}^*) - (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}, \sigma}) - \mathcal{E}_{\mathbf{z}}(f_{\rho}^*))| \geq \varepsilon \} \\ & \leq 2\mathcal{N}\left(B_R, \frac{\varepsilon}{4L}\right) \exp\left\{-\frac{3n(3\varepsilon - D(\lambda))}{6(M + 3C_1)}\right\}, \end{aligned}$$

因 $R = \sqrt{\frac{M}{\lambda}}$, 所以

$$\log \mathcal{N}\left(B_R, \frac{\varepsilon}{4L}\right) = \log \mathcal{N}\left(B_1, \frac{\varepsilon\sqrt{\lambda}}{4L\sqrt{M}}\right).$$

结合引理 3.2, 可得

$$\log \mathcal{N}\left(B_1, \frac{\varepsilon\sqrt{\lambda}}{4L\sqrt{M}}\right) \leq C_0 \left(\left(\log \frac{4L\sqrt{M}}{\varepsilon\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right).$$

因此

$$\begin{aligned} & \text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}(f_{\rho}^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}_{\mathbf{z}}(f_{\rho}^*))| \geq \varepsilon \} \\ & \leq 2 \exp \left\{ C_0 \left(\left(\log \frac{4L\sqrt{M}}{\varepsilon\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right) - \frac{n(3\varepsilon - D(\lambda))}{2(M + 3C_1)} \right\}. \end{aligned}$$

证毕. |

定理 3.3 设 $f_{\rho}^* \in \text{Lip}_{C^1}$, 则对任意正整数 m , 和 $0 < \delta \leq 1$, 至少以 $1 - \delta$ 概率成立

$$\begin{aligned} & \mathcal{E}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}(f_{\rho}^*) \\ & \leq \frac{2(3C_1 + 4M)}{3n} \left\{ C_0 \left(\left(\log \frac{12nL4^{d+1}\sqrt{M}}{2(3C_1 + 4M)C_0m^{d+1}\sqrt{\lambda}} \right)^{d+1} + \frac{m^{d+1}}{4^{d+1}} \right) + 2 \log \frac{2}{\delta} \right\} \\ & \quad + \frac{15dC + 12dM + 39}{\sqrt{m}} + 27M^4\lambda \left(\frac{9m}{4\pi} \right)^{2d}. \end{aligned}$$

其中 C 为 Lipschitz 常数, C_0 和 C_1 分别由引理 3.2 和假定给出.

证 将残差分解为

$$\begin{aligned} \mathcal{E}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}(f_{\rho}^*) & \leq |\mathcal{E}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}(f_{\rho}^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}_{\mathbf{z}}(f_{\rho}^*))| \\ & \quad + |\mathcal{E}_{\mathbf{z}}(Q_d(f_{\rho}^*, \cdot)) - \mathcal{E}_{\mathbf{z}}(f_{\rho}^*) - (\mathcal{E}(Q_d(f_{\rho}^*, \cdot)) - \mathcal{E}(f_{\rho}^*))| + D(\lambda) \\ & = T_1 + T_2 + D(\lambda). \end{aligned}$$

为给出 T_1 的上界, 需要分 $\varepsilon \geq \frac{M}{3n}$ 和 $\varepsilon < \frac{M}{3n}$ 两种情况进行讨论.

(i) 当 $\varepsilon \geq \frac{M}{3n}$, 由定理 3.1 知

$$\begin{aligned} & \text{Prob}_{\mathbf{z} \in Z} \{ |\mathcal{E}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}(f_{\rho}^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}, \sigma})) - \mathcal{E}_{\mathbf{z}}(f_{\rho}^*))| \leq \varepsilon \} \\ & \geq 1 - 2 \exp \left\{ C_0 \left(\left(\log \frac{4L\sqrt{M}}{\varepsilon\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right) - \frac{3n(3\varepsilon - D(\lambda))}{6(M + 3C_1)} \right\} \\ & \geq 1 - 2 \exp \left\{ C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right) - \frac{3n(3\varepsilon - D(\lambda))}{6(M + 3C_1)} \right\}. \end{aligned}$$

设

$$2 \exp \left\{ C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right) - \frac{n(3\varepsilon - D(\lambda))}{2(M + 3C_1)} \right\} = \frac{\delta}{2}.$$

则有

$$\varepsilon = \frac{2(M + 3C_1)C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right)}{3n} + \frac{2(M + 3C_1) \log \frac{4}{\delta}}{3n} + \frac{D(\lambda)}{3} \geq \frac{M}{3n}.$$

则不等式

$$\begin{aligned} & |\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \\ & \leq \frac{2(M + 3C_1)C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right)}{3n} + \frac{2(M + 3C_1) \log \frac{4}{\delta}}{3n} + \frac{D(\lambda)}{3} \end{aligned}$$

至少以 $1 - \frac{\delta}{2}$ 的概率成立.

(ii) 当 $\varepsilon \leq \frac{M}{3n}$, 由定理 3.2 知, 不等式

$$|\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \leq \frac{M}{3n}$$

至少以 $1 - \frac{\delta}{2}$ 的概率成立.

结合 $\varepsilon > \frac{M^2}{3n}$ 和 $\varepsilon \leq \frac{M^2}{3n}$ 这两种情形, 不等式

$$\begin{aligned} T_1 &= |\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}_{\mathbf{z}}(f_\rho^*))| \\ &\leq \frac{2(M + 3C_1)C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{1}{\sigma^{2(d+1)}} \right)}{3n} + \frac{2(M + 3C_1) \log \frac{4}{\delta}}{3n} + \frac{D(\lambda)}{3} \end{aligned}$$

至少以 $1 - \frac{\delta}{2}$ 的概率成立.

由定理 3.1 知, 不等式

$$T_2 = |\{\mathcal{E}_{\mathbf{z}}(G_m(f_\rho^*)) - \mathcal{E}_{\mathbf{z}}(f_\rho^*)\} - \{\mathcal{E}(G_m(f_\rho^*)) - \mathcal{E}(f_\rho^*)\}| \leq \frac{6C_1 + 2M}{3n} \log \frac{4}{\delta} + \frac{1}{2}D(\lambda)$$

至少以 $1 - \frac{\delta}{2}$ 的概率成立.

因此不等式

$$\begin{aligned} & \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) \leq 2\varepsilon_R + \frac{4(3C_1 + M)}{3n} \log \frac{4}{\delta} + 3D(\lambda) \\ & \leq \frac{2(3C_1 + M)}{3n} \left\{ C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{m^{d+1}}{4^{d+1}} \right) + 2 \log \frac{4}{\delta} \right\} + 2D(\lambda) \end{aligned}$$

至少以 $1 - \delta$ 的概率成立.

由定理 2.2 和 (2.3) 式知

$$D(\lambda) \leq \frac{5dC + 4dM + 13}{\sqrt{m}} L + 9M^4 \lambda \left(\frac{9m}{4\pi} \right)^{2d}.$$

因此

$$\begin{aligned} \mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) &\leq \frac{2(3C_1 + M)}{3n} \left\{ C_0 \left(\left(\log \frac{12Ln\sqrt{M}}{M\sqrt{\lambda}} \right)^{d+1} + \frac{m^{d+1}}{4^{d+1}} \right) + 2 \log \frac{4}{\delta} \right\} \\ &\quad + \frac{10dC + 8dM + 26}{\sqrt{m}} L + 18M^4 \lambda \left(\frac{9m}{4\pi} \right)^{2d}. \end{aligned}$$

证毕. |

根据定理 3.3, 我们可直接得到学习算法的泛化误差上界.

定理 3.4 假定 $\sigma = \frac{2}{\sqrt{m}}, m = n^{\frac{2}{2d+3}}, \lambda = n^{-\frac{4d+1}{2d+3}}, f_\rho^* \in \text{Lip}_C 1$. 则对任意正整数 n 和 $0 < \delta < 1$, 下式至少以 $1 - \delta$ 的概率成立

$$\mathcal{E}(\pi_M(f_{\mathbf{z},\sigma})) - \mathcal{E}(f_\rho^*) \leq c_1 n^{-\frac{1}{2d+3}} + c_2 \frac{\log \frac{4}{\delta}}{n} + c_3 \frac{(c_4 + 2 \log n)^{d+1}}{n},$$

其中 $c_1 = \frac{(3C_1+M)C_0}{6 \times 4^d} + 10dCL + 8dML + 26L + \frac{729}{8}M^4, c_2 = \frac{4(3C_1+M)}{3}, c_3 = \frac{2(3C_1+M)C_0}{3}, c_4 = \log \frac{12L}{\sqrt{M}}$.

显然, 参数 σ 和 λ 在定理 3.4 中扮演着关键的角色. 其结果表明, 选择适当的 σ 和 λ , 正则化学习算法具有较好的泛化性能.

4 结论

本文研究在广义凸损失函数和变高斯核情形下正则化学习算法的泛化误差估计问题. 泛化误差被分解为样本误差和正则误差. 利用高斯核函数的特性, 通过直接构建高斯 RBF 神经网络给出正则误差的上界, 通过投影算子和高斯再生核希尔伯特空间的覆盖数给出了样本误差的上界. 在收敛性分析中, 广义损失函数的凸性是非常重要的. 结论表明, 选取适当的 σ 和 λ , 正则化学习算法可获得较好的泛化性能.

参 考 文 献

- [1] Vapnik V, Chervonenkis A. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 1971, **16**: 264-280
- [2] Wu Q, Ying Y M, Zhou D X. Learning rates of least-square regularized regression. *Found Comput Math*, 2006, **6**: 171-192
- [3] Cucker F, Smale S. On the mathematical foundations of learning. *Bull Amer Math Soc*, 2001, **39**: 1-49
- [4] Caponnetto A, DeVito E. Optimal rates for the regularized least-squares algorithm. *Found Comput Math*, 2007, **7**: 331-368
- [5] Cybenko G. Approximation by superposition of sigmodal functions. *Mathematics of Control, Signal, and Systems*, 1989, **3**: 303-314
- [6] Esposito A, Marinaro M, Oricchio D, Scarpetta S. Approximation of continuous and discontinuous mappings by a growing neural RBF-based algorithm. *Neural Networks*, 2000, **13**: 651-665
- [7] Smale S, Zhou D X. Estimating the approximation error in learning theory. *Anal Appl*, 2003, **1**: 17-41
- [8] Smale S, Zhou D X. Shannon sampling and function reconstruction from point values. *Bull Amer Math Soc*, 2004, **41**: 279-305
- [9] Tong H Z, Chen D R, Li Z P. Learning rates for regularized classifiers using multivariate polynomial kernel. *J Complexity*, 2008, **24**: 619-631
- [10] Wu Q, Ying Y M, Zhou D X. Learning theory: from regression to classification. *Studies in Computational Mathematics*, 2006, **12**: 257-290

- [11] Zhou D X, Jetter K. Approximation with polynomial kernels and SVM classifiers. *Adv Comput Math*, 2006, **25**: 323–344
- [12] Aronszajn N. Theory of reproducing kernels. *Trans Amer Soc*, 1950, **68**: 337–404
- [13] Xiang D H, Zhou D X. Classification with gaussians and convex loss. *J Mach Learning Res*, 2009, **10**: 1447–1468
- [14] Steinwart I. Support vector machines are universally consistent. *J Complexity*, 2002, **18**: 768–791
- [15] Ye G B, Zhou D X. Learning and approximation by Gaussians on Riemannian manifolds. *Adv Comput Math*, 2008, **29**: 291–310
- [16] Maierov V. Approximation by neural networks and learning theory. *J Complexity*, 2006, **22**: 102–117
- [17] Wu Q, Zhou D X. Support vector machine classifiers: linear programming versus quadratic programming. *Neural Computation*, 2005, **17**: 1160–1187
- [18] Xie T F, Cao F L. The rate of approximation of Gaussian radial basis neural networks in continuous function space. *Acta Mathematica Sinica*, 2013, **29**: 295–302
- [19] Cucker F, Zhou D X. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge: Cambridge University Press, 2007
- [20] Bartlett P L. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of network. *IEEE Trans Inform Theory*, 1998, **44**: 525–536
- [21] Zhou D X. The covering number in learning theory. *J Complexity*, 2002, **18**: 739–767
- [22] Zhou D X. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans Inform Theory*, 2003, **49**: 1734–1752
- [23] Williamson R C, Smola A J, Schölkopf B. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Trans Inform Theory*, 2001, **47**: 2516–2532
- [24] Pontil M. A note different covering numbers in learning theory. *J Complexity*, 2003, **19**: 665–671

Generalization Bounds of Compressed Regression Learning Algorithm

^{1,2}Zhang Yongquan ¹Li Youmei

(¹*Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou 310018;*

²*Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049)*

Abstract: This article studies generalization performance of regularized learning algorithm with a general convex loss function and varying Gaussian kernels. Our goal is to give a satisfactory estimate of generalization error for the learning algorithm. The generalization error is measured by regularization error and sample error. The regularization error is estimated by constructing a radial basis function (briefly denoted by RBF) neural network in view of the special structures of Gaussian kernels. The sample error is obtained by using projection operator and covering number of reproducing kernel Hilbert spaces with Gaussian kernels. The obtained results demonstrate the learning algorithm has good generalization performance with suitable choice of m and λ .

Key words: Learning theory; RBF neural network; Gaussian kernels; Generalization error.

MR(2000) Subject Classification: 41A25; 41A63; 42B08