中国科学: 信息科学 2011 年 第 41 卷 第 2 期: 144–156
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
论 文

多元回归学习算法收敛速度的估计

徐宗本[①], 张永全[①], 曹飞龙[②]*

① 西安交通大学信息与系统科学研究所, 西安 710049
② 中国计量学院计量与计算科学研究, 杭州 310018
* 通信作者. E-mail: flcao@263.net

收稿日期: 2009–12–04; 接受日期: 2010–03–05
国家重点基础研究发展计划 (批准号: 2007CB311002) 和国家自然科学基金 (批准号: 90818020, 60873206) 资助项目

摘要 在许多应用中, 回归函数的先验信息往往不能事先获取. 因此, 有必要利用有效的方法学习回归函数. 本文研究学习理论中的回归问题, 即研究多项式空间上具有最小二乘平方损失正则学习算法的收敛速度问题. 主要目的在于分析学习理论中多维回归问题的泛化误差. 利用逼近论中著名 Jackson 算子、覆盖数理论、集合的熵数以及有关概率不等式, 得到学习算法收敛速度的上、下界估计. 特别地, 对于满足一定条件的多元光滑回归函数, 除一个对数因子外, 所获的收敛速度是最优的. 本文结果对研究回归学习算法的收敛性、稳定性及复杂性等有着重要的意义.

关键词 学习理论 收敛速度 覆盖数 熵数

1 引言

周知, 回归与分类是学习理论中两个重要的基本问题. 回归问题一直是学习理论研究的热点之一, 而对回归问题收敛性的研究又是近几年人们关注的核心问题之一. 早在 1998 年, Vapnik[1] 利用 VC 维较系统地研究了统计学习理论中分类和回归算法的收敛性问题, 同时, 对结构风险极小化原则算法的相容性进行了研究, 得出了一系列相关的收敛性结果. 在文献 [2] 中, Shawe-Taylor 等利用 VC 维研究了结构风险极小化算法的泛化性能. 2001 年, Cucker 与 Smale[3] 从数学的观点系统地给出了学习理论的基础框架, 并指出回归问题是学习理论中的重要问题之一, 他们进而利用覆盖数方法研究了学习理论中的逼近问题. 2002 年, Cucker 与 Smale[4] 进一步就如何构造回归函数的近似函数以及选择正则参数使得具有较好逼近阶等问题进行了研究. 2006 年, Wu 等 [5] 利用概率论中的经典 Bernstien 不等式在再生核 Hilbert 空间中研究最小二乘损失正则回归学习算法的收敛性问题. 2007 年, Smale 等 [6] 利用再生核的积分算子研究了回归学习算法的逼近问题; Caponnetto 与 DeVito[7] 对再生核 Hilbert 空间中正则最小二乘回归算法的性能进行了分析, 特别地, 对向量值目标函数给出了相应的收敛速度估计. 在文献 [8] 中, Temlyakov 利用函数逼近论中的方法和技巧研究了学习算法的逼近问题. 有关学习算法收敛速度的研究还可以见新近文献 [9–11].

然而, 我们注意到, 已有的关于学习算法收敛速度的研究基本上限于学习算法收敛速度的上界估计. 明显地，仅有上界估计对于刻画学习算法的本质特性是远远不够的, 因为我们无法判断所给出的上界是否最小. 因此, 为了准确地反映学习算法的本质收敛速度, 不仅需要对学习算法收敛速度的上

引用格式: 徐宗本, 张永全, 曹飞龙. 多元回归学习算法收敛速度的估计. 中国科学: 信息科学, 2011, 41: 144–156

界进行估计, 而且更希望对其下界作出估计. 自然地，这样的下界估计是比较困难的, 但却是非常重要而有意义的. 在文献 [12] 中, Temlyakov 利用逼近论的最佳逼近理论研究了学习算法的收敛速度问题, 并通过对回归函数的假定, 给出其收敛速度的一个下界估计. 最近, Steinwart 等 [13] 把相应积分算子的特征值作为复杂性度量研究了学习速度的上、下界.

上述提及的一系列关于学习算法收敛性与收敛速度的研究, 特别是学习算法收敛速度的下界估计, 对刻画学习算法的性能与算法的稳定性、复杂性研究起着十分重要的作用.

然而, 对于一般核所对应的积分算子, 计算它的特征值是非常困难的. 另一方面, 我们知道, 覆盖数在学习理论研究中通常作为复杂性度量而被广泛使用 (见文献 [14–18]). 因此, 本文先利用再生核 Hilbert 空间的覆盖数作为度量估计学习速度的上界. 然后, 我们引进集合的熵数并以此为工具研究学习速度的下界. 我们所获得的上、下界估计除一个对数因子外, 它们的阶是相同的.

2 回归学习算法及其收敛速度的上界

令 $\mathcal{R}^s$ 为 $s(s$ 是正整数) 维欧几里德空间, $[-1,1]^s$ 为 $\mathcal{R}^s$ 中的立方体. 本文以 $X=[-1,1]^s$ 上多项式函数全体所组成的集合作为假设空间, 并在此空间上极小化最小二乘风险. 在回归分析中, 主要考虑满足 $E\mathcal{Y}^2<\infty(\mathcal{Y}^2$ 的数学期望) 的 $\mathcal{R}^s\times\mathcal{R}$ 值随机变量 $(\mathcal{X},\mathcal{Y})$ 以及 $\mathcal{X}$ 的函数值对 $\mathcal{Y}$ 的依赖性. 即, 学习的目标为寻找一个函数 $f:\mathcal{R}^s\to\mathcal{R}$ 使得 $f(\mathcal{X})$ 能够较好地逼近 $\mathcal{Y}$. 我们研究的主要目的是极小化预测误差平方的期望或 L_2 风险 $\mathcal{E}(f)=\mathrm{E}\{|f(\boldsymbol{x})-y|^2\}$. 称使得上述误差达到最小的函数为回归函数, 其定义为

$$m(\boldsymbol{x})=\mathrm{E}\{\mathcal{Y}|\mathcal{X}=\boldsymbol{x}\},\quad \boldsymbol{x}\in\mathcal{R}^s.$$

令 $f:\mathcal{R}^s\to\mathcal{R}$ 为 $\mathcal{R}^s$ 上任意一个可测函数, 用 ν 表示 X 上的分布. 熟知

$$\mathrm{E}\{|f(\boldsymbol{x})-y|^2\}=\mathrm{E}\{|m(\boldsymbol{x})-y|^2\}+\int_{\mathcal{R}^s}(f(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x}).$$

根据 L_2 风险极小化 [19,20] 可知, 回归函数是最好的预测函数, 即

$$\mathrm{E}\{|m(\boldsymbol{x})-y|^2\}=\min_{f:\mathcal{R}^s\to\mathcal{R}}E\{|f(\boldsymbol{x})-y|^2\}.$$

由此可以看出, 当且仅当

$$\mathcal{E}(f)=\mathrm{E}\{|f(\boldsymbol{x})-y|^2\}\tag{1}$$

较小时, 函数 f 为回归函数较好的预测函数. 为说明预测的好坏, 有必要度量函数 f 代替回归函数所产生的误差 (1).

在应用中, 样本的分布往往是未知的. 因而, 回归函数也通常是未知的. 但是, 样本可以根据同一分布采取, 这就导致所谓的回归估计问题. 令 $\boldsymbol{z}=\{z_i\}_{i=1}^n=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$ 为 $X\times Y$ 上独立同分布的样本点集合, 我们的目标是构造回归函数的一个估计子 $f_{\boldsymbol{z}}(\cdot)$, 使得 L_2 误差 $\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x})$ 较小.

假定存在某正数 M, 使得 $|m(\boldsymbol{x})|\leqslant M/4$. 同时, 我们需要对回归函数作一些光滑性假定. 这种假定是合理的, 因为学习的过程往往不是凭空发生的, 总是在一定的环境下产生, 需要一定的学习框架, 即, 假设空间. 在学习理论的研究中, 常取函数空间的形式, 如多项式空间、连续函数空间等 [5,19]. 我们研究的目的就是在这些假设空间上寻找回归函数的最佳逼近.

下面介绍 $X=[-1,1]^s$ 上的多项式函数 [20]. 令 $\mathcal{H}_d$ 为如下函数的集合:

$$f:\mathcal{R}^s\to\mathcal{R},\ f(\boldsymbol{x})=\sum_{0\leqslant k_1\leqslant d}\cdots\sum_{0\leqslant k_s\leqslant d}a_{k_1,\ldots,k_s}x_1^{k_1}x_2^{k_2}\cdots x_s^{k_s},\boldsymbol{x}\in[-1,1]^s,$$

其中对于所有的 $|k_1|\leqslant d,\ldots,|k_s|\leqslant d$, $a_{k_1,\ldots,k_s}\in\mathcal{R}$. 从集合 $\mathcal{H}_d$ 的定义可以看出, 多项式空间 $\mathcal{H}_d$ 的维数 $\dim\mathcal{H}_d=(2d)^s$.

本文考虑以集合 $\mathcal{F}_d=\{f\in\mathcal{H}_d:|f(\boldsymbol{x})|\leqslant M/4,\boldsymbol{x}\in[-1,1]^s\}$ 作为假设空间, 定义估计子:

$$f_{\boldsymbol{z}}=\arg\min_{f\in\mathcal{F}_d}\frac{1}{n}\sum_{i=1}^n(f(\boldsymbol{x}_i)-y_i)^2,\tag{2}$$

其中 $f(\boldsymbol{x})=\sum_{0\leqslant k_1\leqslant d}\cdots\sum_{0\leqslant k_s\leqslant d}a_{k_1,\ldots,k_s}x_1^{k_1}x_2^{k_2}\cdots x_s^{k_s},\boldsymbol{x}\in[-1,1]^s$.

下面, 我们将分析函数 $f_{\boldsymbol{z}}$ 对回归函数 m 的收敛速度. 事实上, $f_{\boldsymbol{z}}$ 与 m 的误差对算法 (2) 的有效性起到关键作用. 由回归函数 m 的定义知

$$\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x})=\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m).$$

我们的目的是在函数集合 $\mathcal{F}_d$ 上估计上述误差的大小.

设经验风险为 $\mathcal{E}_{\boldsymbol{z}}(f)=\frac{1}{n}\sum_{i=1}^n(f(\boldsymbol{x}_i)-y_i)^2$. 事实上, 它是期望风险 $\mathcal{E}(f)$ 的离散形式. 因此, 函数 $f_{\boldsymbol{z}}$ 可以写成 $f_{\boldsymbol{z}}=\arg\min_{f\in\mathcal{F}_d}\mathcal{E}_{\boldsymbol{z}}(f)$.

以下定理 1 给出 L_2 误差的一个上界估计.

定理 1　令集合 $\mathcal{F}_d$ 为假设空间, 那么对于 (2) 式所定义的估计子 $f_{\boldsymbol{z}}$, 不等式

$$\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\leqslant\frac{204M^2}{n}\log\frac{2}{\delta}+\frac{64M^2(2d)^s\log(4M^2n)}{n}+3(\mathcal{E}(Q_d(m))-\mathcal{E}(m))$$

至少以 $1-\delta$ 的置信成立.

若对回归函数加一些光滑性限制, 则从定理 1 可导出下面推论 1 的收敛速度估计. 先给出函数连续模的定义. 记 $[-1,1]^s$ 上的连续函数全体为 $C_{[-1,1]^s}$. 对于 $f\in C_{[-1,1]^s}$, 以及正整数 r, 定义函数 f 的差分 $\Delta_{\boldsymbol{t}}^1f(\boldsymbol{x})=f(\boldsymbol{x}+\boldsymbol{t})-f(\boldsymbol{x})$, $\Delta_{\boldsymbol{t}}^rf(\boldsymbol{x})=\Delta_{\boldsymbol{t}}\Delta_{\boldsymbol{t}}^{r-1}f(\boldsymbol{x})$. 显然有

$$\Delta_{\boldsymbol{t}}^rf(\boldsymbol{x})=\sum_{j=0}^r(-1)^j\binom{r}{j}f(\boldsymbol{x}+j\boldsymbol{t}).$$

令 $\|\boldsymbol{t}\|_2=(t_1^2+\cdots+t_s^2)^{1/2}$ 为向量 $\boldsymbol{t}=(t_1,\ldots,t_s)$ 的欧氏范数, 那么函数 f 的 r 阶连续模定义为

$$\omega_r(f,h)=\sup_{\|\boldsymbol{t}\|_2\leqslant h}\|\Delta_{\boldsymbol{t}}f\|_\infty.$$

这里 $\|f\|_\infty=\max_{\boldsymbol{x}}\in[-1,1]^s|f(\boldsymbol{x})|$.

连续模是逼近论中常用的工具, 通常用来刻画函数光滑性与逼近误差的工具, 它具有如下的性质.

命题 1[21]　如下论断成立:

(1) 对于 $\lambda>0$, 有 $\omega_r(f,\lambda h)\leqslant(1+\lambda)^r\omega_r(f,h)$.

(2) 如果函数 f 关于变量 t_i 的偏导数用 ∂_i 表示, 以及 $\partial^k=\partial_1^{k_1}\cdots\partial_s^{k_s}$, 则

$$\omega_r(f,h)\leqslant h^r\sum_{k_1+\cdots+k_s=r}\frac{r!}{k_1!\cdots k_s!}\|\partial^kf\|_\infty.$$

下面, 我们给出函数偏导数属于 Lipschitz 类的定义.

定义 1 假定 k 为自然数, $0<\beta\leqslant 1$ 并令 $C>0$. $\alpha_i(i=1,2,\ldots,s)$ 为非负整数, 向量 $\alpha=(\alpha_1,\ldots,\alpha_s)$, 且 $\sum_{j=1}^{s}\alpha_j=k$, 函数 f 定义在 $[-1,1]^s$ 上且其偏导数 $\frac{\partial^k f}{\partial x_1^{\alpha_1}\cdots\partial x_s^{\alpha_s}}$ 存在. 如果对所有 $\boldsymbol{x},\boldsymbol{z}\in[-1,1]^s$, 存在常数 $C>0$ 使得

$$\left|\frac{\partial^k f(\boldsymbol{x})}{\partial x_1^{\alpha_1}\cdots\partial x_s^{\alpha_s}}-\frac{\partial^k f(\boldsymbol{z})}{\partial z_1^{\alpha_1}\cdots\partial z_s^{\alpha_s}}\right|\leqslant C\|\boldsymbol{x}-\boldsymbol{z}\|^{\beta},$$

则称函数 f 的 k 阶偏导数属于 $\mathrm{Lip}_C\beta$ 类, 记作 $\partial^k f\in\mathrm{Lip}_C\beta$.

根据上述的定义, 我们从定理 1 得到如下的推论 1.

推论 1 令集合 $\mathcal{F}_d$ 为假设空间, 假定 k 为自然数, $0<\beta\leqslant 1$. 如果回归函数 $m(\boldsymbol{x})$ 的 k 阶偏导数属于 $\mathrm{Lip}_C 1$, 取 $d=[n^{1/(k+1)}]$, 则有

$$E\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x})\leqslant\frac{408M^2\log 2}{n}+2C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

其中 $C_{K,s}=2(108M^2 2^s)^{\frac{2k}{2k+s}}+2(3C_K^{'})^{\frac{s}{2k+s}}$, $[a]$ 表示实数 a 的整数部分.

为证明上面两个结论, 我们需要介绍下 Jackson 算子, 周知, 它在逼近论中具有极为重要的作用 [21]. 对于自然数 d,r, 令 $q=[d/r]+1$, 一元 Jackson 核函数为

$$K_{dr}(t)=L_{q,r}(t)=\frac{1}{\lambda_{qr}}\left(\frac{\sin\frac{qt}{2}}{\sin\frac{t}{2}}\right)^{2r},\tag{3}$$

其中 $\lambda_{qr}=\int_{-\pi}^{\pi}(\frac{\sin\frac{qt}{2}}{\sin\frac{t}{2}})^{2r}\mathrm{d}t$.

熟知, Jackson 核函数具有如下的性质.

引理 1[20] 令 Jackson 核函数如 (3) 式所定义. 那么 $K_{dr}(t)$ 为一个 d 阶三角多项式, 并有

$$\int_{-\pi}^{\pi}K_{dr}(t)\mathrm{d}t=1,\quad\int_{-\pi}^{\pi}t^k K_{dr}(t)\mathrm{d}t\leqslant C_k(d+1)^{-k},\ k=0,1,\ldots,2r-2.$$

令 $K_{dr}(\boldsymbol{t})=K_{dr}(t_1)\cdots K_{dr}(t_s)$ 并记 $|k|=k_1+\cdots+k_s$, 则不难得到

$$\int_{[-\pi,\pi]^s}\boldsymbol{x}^k K_{dr}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}\leqslant C_k'(d+1)^{-k_1}\cdots(d+1)^{-k_s}=C_k'(d+1)^{-|k|},$$

其中 $C_k'=C_{k_1}\cdots C_{k_s}$.

令 $\phi(\boldsymbol{t})=f(\cos\boldsymbol{t})$, 则 $\phi(\boldsymbol{x})$ 为周期函数. 在 $[-\pi,\pi]^s$ 上定义 Jackson 型算子:

$$\begin{aligned}J_d(f,\boldsymbol{u})=&-\int_{[-\pi,\pi]^s}K_{dr}(\boldsymbol{t})\sum_{j=1}^{k}(-1)^j\binom{k}{j}f(\cos(\boldsymbol{u}+j\boldsymbol{t}))\mathrm{d}\boldsymbol{t}\\=&-\int_{[-\pi,\pi]^s}K_{dr}(\boldsymbol{t})\sum_{j=1}^{k}(-1)^j\binom{k}{j}\phi(\boldsymbol{u}+j\boldsymbol{t})\mathrm{d}\boldsymbol{t},\end{aligned}\tag{4}$$

其中 r 为满足 $r\leqslant[(k+2)/2]$ 的最小正整数.

$L^2_{[-\pi,\pi]^s}$ 表示 $[-\pi,\pi]^s$ 上的 Lebesgue 平方可积函数全体, 如果赋予范数 $\|f\|_2=(\int_{[-\pi,\pi]^s}|f(\boldsymbol{t})|^2\mathrm{d}\boldsymbol{t})^{1/2}<\infty$, 则 $L^2_{[-\pi,\pi]^s}$ 是一 Banach 空间. 如果在 $L^2_{[-\pi,\pi]^s}$ 上定义内积 $\langle f,g\rangle=\int_{[-\pi,\pi]^s}f(\boldsymbol{x})g(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$,

则 $L^2_{[-\pi,\pi]^s}$ 是一 Hilbert 空间, 并且 $\{e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\}_{k\in\mathbb{Z}^s}$ 为 Hilbert 空间 $L^2_{[-\pi,\pi]^s}$ 关于上述内积的规范正交基, 则对于任意的 $\psi\in L^2_{[-\pi,\pi]^s}$, 有 $\psi(\boldsymbol{x})=\sum_{k\in\mathbb{Z}^s}a_{\boldsymbol{k}}(\psi)e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, 其中 $a_{\boldsymbol{k}}(\psi)=\int_{[-\pi,\pi]^s}\psi(\boldsymbol{t})e^{i\boldsymbol{k}\cdot\boldsymbol{t}}\mathrm{d}\boldsymbol{t}$, $\mathbb{Z}^s$ 表示 s 重整数指标集.

下面约定向量 $\boldsymbol{z}$ 被整数 j 整除当且仅当向量的每个分量都能被 j 整除.

引理 2 令 $\phi\in L^2_{[-\pi,\pi]^s}$, $\boldsymbol{l}=(l_1,\ldots,l_s)$, l_j 是正整数 ($j=1,2,\ldots$), 则当 $\boldsymbol{l}$ 不被 j 整除时, 则有 $\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}\mathrm{d}\boldsymbol{t}=0$.

证明 对于函数 $\phi\in L^2_{[-\pi,\pi]^s}$, 函数 $\phi(j\boldsymbol{t})$ 关于每个变量具有周期 $2\pi/j$, 则有

$$\begin{aligned}
&\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}\mathrm{d}\boldsymbol{t}=\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}\phi(jt_1,\ldots,jt_s)e^{i(l_1t_1+\cdots+l_st_s)}\mathrm{d}t_1\cdots\mathrm{d}t_s\\
=&\int_{-\pi+2il_1\pi/j}^{\pi+2il_1\pi/j}\cdots\int_{-\pi}^{\pi}\phi(jt_1,\ldots,jt_s)e^{i(l_1t_1+\cdots+l_st_s)}\mathrm{d}t_1\cdots\mathrm{d}t_s\\
=&\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}\phi(jt_1,\ldots,jt_s)e^{i((l_1t_1+2il_1\pi/j)+\cdots+l_st_s)}\mathrm{d}t_1\cdots\mathrm{d}t_s\\
=&e^{2il_1\pi/j}\int_{[-\pi,\pi]}\cdots\int_{[-\pi,\pi]}\phi(jt_1,\ldots,jt_s)e^{i(l_1t_1+\cdots+l_st_s)}\mathrm{d}t_1\cdots\mathrm{d}t_s\\
=&e^{2il_1\pi/j}\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}\mathrm{d}\boldsymbol{t}.
\end{aligned}$$

因此, 当 l_1 不被 j 整除时, 有

$$\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}\mathrm{d}\boldsymbol{t}=0.$$

同理可证: 当 $\boldsymbol{l}$ 的其他分量 $l_i,i=2,3,\ldots,s$ 不能被 j 整除时, 上述积分均为 0. 所以, 当 $\boldsymbol{l}$ 不被 j 整除时, 就有

$$\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})e^{i\boldsymbol{l}\cdot\boldsymbol{t}}\mathrm{d}\boldsymbol{t}=0.$$

引理 2 得证.

根据引理 2 及以下事实

$$\begin{aligned}
\int_{[-\pi,\pi]^s}\phi(\boldsymbol{u}+j\boldsymbol{t})\cos\boldsymbol{l}\cdot\boldsymbol{t}\mathrm{d}\boldsymbol{t}&=\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})\cos\boldsymbol{l}\cdot\left(\boldsymbol{t}-\frac{\boldsymbol{k}}{j}\right)\mathrm{d}\boldsymbol{t}\\
&=\frac{1}{2}\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})\left(e^{i\boldsymbol{l}\cdot(\boldsymbol{t}-\frac{\boldsymbol{k}}{j})}+e^{-i\boldsymbol{l}\cdot(\boldsymbol{t}-\frac{\boldsymbol{k}}{j})}\right)\mathrm{d}\boldsymbol{t}\\
&=\frac{1}{2}\int_{[-\pi,\pi]^s}\phi(j\boldsymbol{t})(e^{i\boldsymbol{k}\cdot\boldsymbol{t}}e^{-i\boldsymbol{l}\cdot\frac{\boldsymbol{k}}{j}}+e^{-i\boldsymbol{k}\cdot\boldsymbol{t}}e^{i\boldsymbol{l}\cdot\frac{\boldsymbol{k}}{j}})\mathrm{d}\boldsymbol{t},
\end{aligned}\tag{5}$$

其中 $\cos\boldsymbol{l}\cdot\boldsymbol{t}=\cos l_1t_1\cdots\cos l_st_s$. 则当 $\boldsymbol{l}$ 不能被 j 所整除时, (5) 式为零. 否则, 对于 $1\leqslant j\leqslant k$, $0\leqslant l_i\leqslant d(i=1,2,\ldots,d)$, (5) 式为一个三角多项式. 从上述的讨论可知, K_{dr} 关于每个变量均为 d 阶三角多项式. 而对于 $1\leqslant j\leqslant k,0\leqslant l_i\leqslant d(i=1,2,\ldots,d)$, $J_d(f,\boldsymbol{u})$ 为 $\int_{-\pi}^{\pi}f(\boldsymbol{u}+j\boldsymbol{t})\cos\boldsymbol{l}\cdot\boldsymbol{t}\mathrm{d}\boldsymbol{t}$ 的一个线性组合. 所以, $J_d(f,\boldsymbol{u})$ 为一个三角多项式. 于是令 $\boldsymbol{u}=\arccos\boldsymbol{x}=\arccos x_1\cdots\arccos x_s$, 则 $Q_d(f,\boldsymbol{x})=J_d(f,\arccos\boldsymbol{x})$ 为 $2d-2$ 次代数多项式.

命题 2 令 k 为一个自然数, $f\in C_{[-1,1]^s}$. 对于 $d=0,1,\ldots$, 成立

$$|f(\boldsymbol{x})-Q_d(f,\boldsymbol{x})|\leqslant C_{ks}\omega_k\left(f,\frac{1}{d+1}\right),\quad\forall\boldsymbol{x}\in[-1,1]^s.$$

证明 根据引理 1、$Q_d(f,\boldsymbol{x})$ 的定义以及 $K_{dr}(\boldsymbol{t})=K_{dr}(-\boldsymbol{t})$, 并令 $\boldsymbol{x}=\cos\boldsymbol{u}$, 则有

$$\begin{aligned}|f(\boldsymbol{x})-Q_d(f,\boldsymbol{x})| &= |f(\cos\boldsymbol{u})-J_d(f,\boldsymbol{u})| = \left|\int_{[-\pi,\pi]^s} K_{dr}(\boldsymbol{t})\triangle^k_{\|\boldsymbol{t}\|_2}\phi(\boldsymbol{u})d\boldsymbol{t}\right| \\ &\leqslant \int_{[-\pi,\pi]^s} K_{dr}(\boldsymbol{t})|\triangle^k_{\|\boldsymbol{t}\|_2}\phi(\boldsymbol{u})|\mathrm{d}\boldsymbol{t} \leqslant 2^s\int_{[0,\pi]^s} K_{dr}(\boldsymbol{t})\omega_k(\phi,\|\boldsymbol{t}\|_2)\mathrm{d}\boldsymbol{t},\end{aligned}$$

其中 $\phi(\boldsymbol{u})=f(\cos\boldsymbol{u})$.

对于任何 $\boldsymbol{t},\boldsymbol{t}'\in\mathcal{R}^d$, 都有 $|\cos y-\cos y'|\leqslant|y-y'|$. 从而

$$\begin{aligned}\|\boldsymbol{t}-\boldsymbol{t}'\|_2 &= \sqrt{|t_1-t_1'|^2+\cdots+|t_s-t_s'|^2} \\ &\geqslant \sqrt{|\cos t_1-\cos t_1'|^2+\cdots+|\cos t_s-\cos t_s'|^2} = \|\cos\boldsymbol{t}-\cos\boldsymbol{t}'\|_2.\end{aligned}$$

所以

$$\begin{aligned}\sup_{\|\boldsymbol{t}-\boldsymbol{t}'\|_2\leqslant h}|\phi(\boldsymbol{t})-\phi(\boldsymbol{t}')| &\leqslant \sup_{\|\cos\boldsymbol{t}-\cos\boldsymbol{t}'\|_2\leqslant h}|\phi(\boldsymbol{t})-\phi(\boldsymbol{t}')| \\ &= \sup_{\|\boldsymbol{x}-\boldsymbol{t}'\|_2\leqslant h}|f(\boldsymbol{x})-f(\boldsymbol{x}')|,\end{aligned}$$

即 $\omega_k(\phi,\|\boldsymbol{t}\|_2)\leqslant\omega_k(f,\|\boldsymbol{t}\|_2)$.

由函数 f 的连续模定义, 我们知

$$\omega_k(f,\|\boldsymbol{t}\|_2)\leqslant(1+(d+1)\|\boldsymbol{t}\|_2)^k\omega_k\left(f,\frac{1}{d+1}\right).$$

对于 $n=0,1,\ldots,k\leqslant 2r-2$ 及任意 $\boldsymbol{x}\in[-1,1]^s$, 应用引理 1 可得

$$\begin{aligned}|f(\boldsymbol{x})-Q_d(f,\boldsymbol{x})| &\leqslant 2^s\omega_k\left(f,\frac{1}{d+1}\right)\int_{[0,\pi]^s}(1+(d+1)\|\boldsymbol{t}\|_2)^k K_{dr}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \\ &\leqslant C_{ks}\omega_k\left(f,\frac{1}{d+1}\right),\end{aligned}$$

其中 $C_{ks}=2^sC_k'$. 命题 2 证毕.

为估计误差 $\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)$, 我们需要估计

$$\begin{aligned}\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m) &= \{(\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m))\}+\{\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(Q_d(m))\} \\ &\quad+\{(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))-(\mathcal{E}(Q_d(m))-\mathcal{E}(m))\}+(\mathcal{E}(Q_d(m))-\mathcal{E}(m)) \\ &\leqslant \{(\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m))\}+(\mathcal{E}(Q_d(m))-\mathcal{E}(m)) \\ &\quad+\{(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))-(\mathcal{E}(Q_d(m))-\mathcal{E}(m))\}. \qquad (6)\end{aligned}$$

首先, 我们估计 (6) 式中的第三项, 对于随机变量 $\xi=(Q_d(m,\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2$, 我们需要如下的引理 3.

引理 3[14] 令 ξ 为 $Z=[-1,1]^s\times[-M,M]$ 上的随机变量且其均值、方差分别为 μ,σ^2. 假定 $\mu\geqslant 0,|\xi-\mu|\leqslant B$ 几乎处处成立. 并且 $\mathrm{E}(\xi^2)\leqslant c_\xi\mathrm{E}(\xi)$, 那么对于每个 $\varepsilon>0$, 成立如下不等式

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\left\{\frac{\mu-\frac{1}{n}\sum_{i=1}^n\xi(z_i)}{\sqrt{\mu+\varepsilon}}\geqslant\sqrt{\varepsilon}\right\}\leqslant\exp\left\{-\frac{n\varepsilon}{2c_\xi+\frac{2}{3}B}\right\}.$$

定理 2 对于任意的 $0<\delta\leqslant 1$, 以下不等式至少以 $1-\frac{\delta}{2}$ 置信成立:

$$(\mathcal{E}(Q_d(m))-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))\leqslant\frac{70M^2}{n}\log\frac{2}{\delta}+\frac{1}{2}(\mathcal{E}(Q_d(m))-\mathcal{E}(m)).$$

证明 从 (4) 式知, $Q_d(f,\boldsymbol{x})\in\mathcal{H}_d$. 根据命题 2, 可知

$$|Q_d(f,\boldsymbol{x})|\leqslant|f(\boldsymbol{x})|+\omega_k\left(f,\frac{1}{d}\right).$$

若 $f(\boldsymbol{x})$ 的 k 阶偏导数满足 Lip_C1, 那么

$$\omega_k\left(f,\frac{1}{d}\right)\leqslant C_k'\frac{1}{d^k},$$

其中 C_k' 为依赖于 k 的常数.

当 $d^k\geqslant\frac{C_k'}{M}$ 时, 则有

$$|Q_d(f,\boldsymbol{x})|\leqslant M+C_k^{'}\frac{1}{d^k}\leqslant 2M,\qquad\forall\boldsymbol{x}\in[-1,1]^s.$$

因此, $Q_d(f,\boldsymbol{x})\in\mathcal{F}_d$, 以及 $|m(\boldsymbol{z})|\leqslant M$. 于是可得

$$|\xi|=|(Q_d(m,\boldsymbol{z})-m(\boldsymbol{z}))(Q_d(\boldsymbol{z})+m(\boldsymbol{z})-2y)|\leqslant 15M^2.$$

则有

$$|\xi-\mu|\leqslant B=30M^2.$$

而

$$\sigma^2\leqslant\mathrm{E}(\xi^2)\leqslant c_\xi\mathrm{E}(\xi)=25M^2\mathrm{E}(\xi).$$

把 $(Q_d(m,\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2$ 应用到引理 3, 则不等式

$$\begin{aligned}(\mathcal{E}(Q_d(m))-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))&\leqslant\sqrt{\varepsilon(\mathcal{E}(Q_d(m))-\mathcal{E}(m)+\varepsilon)}\\&\leqslant\varepsilon+\frac{1}{2}(\mathcal{E}(Q_d(m))-\mathcal{E}(m))\end{aligned}$$

至少以 $1-\exp\{-\frac{n\varepsilon}{70M^2}\}$ 置信成立.

令 $\exp\{-\frac{n\varepsilon}{70M^2}\}=\frac{\delta}{2}$, 则有 $\varepsilon=\frac{70M^2}{n}\log\frac{2}{\delta}$. 从而就有

$$(\mathcal{E}(Q_d(m))-\mathcal{E}(m))-(\mathcal{E}_{\boldsymbol{z}}(Q_d(m))-\mathcal{E}_{\boldsymbol{z}}(m))\leqslant\frac{70M^2}{n}\log\frac{2}{\delta}+\frac{1}{2}(\mathcal{E}(Q_d(m))-\mathcal{E}(m))$$

至少以 $1-\frac{\delta}{2}$ 置信成立.

对于 (6) 式的第一部分, 由于随机变量 $\xi=(f_{\boldsymbol{z}}(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2$ 不是 $[-1,1]^s\times[-M,M]$ 上单一的随机变量, 而依赖于样本 $\boldsymbol{z}$. 变量 ξ 在函数集合 $\mathcal{F}_d$ 随着样本的改变而改变, 因此不应把此变量看作一个固定的函数. 下面, 利用单位球的覆盖数估计 (6) 式的第一部分.

定义 2[17] 令 $\mathcal{F}$ 为一度量空间的子集. 对于任意的 $\varepsilon>0$, $\mathcal{F}$ 的覆盖数定义为最小的正整数 l 使得 l 个以 ε 为半径的球覆盖集合 $\mathcal{F}$, 记为 $l=\mathcal{N}(\mathcal{F},\varepsilon)$.

学习理论研究中经常用以覆盖数作为度量 (见文献 [15,16,18,22–25]). 如果 B_R 是 r 维空间中半径为 R 的闭球, 则成立 [18]

$$\log\mathcal{N}(B_R,\varepsilon)\leqslant r\log\frac{4R}{\varepsilon}.\tag{7}$$

为了估计 (6) 式中的 $\{\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\}-\{\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m)\}$, 我们还需要如下的引理 4.

对于 $[-1,1]^s\times[-M,M]$ 上的函数 g, 我们用 $\mathrm{E}(g)=\int_Z g(z)\mathrm{d}\rho$ 表示函数 g 的期望.

引理 4[14] 设 $c_0\geqslant 0,\mathcal{G}$ 为 $Z=[-1,1]^s\times[-M,M]$ 上的一函数集合, 使得对于每个 $g\in\mathcal{G}$, $|g-\mathrm{E}(g)|\leqslant B$, $\mathrm{E}(g^2)\leqslant c_0\mathrm{E}(g)$ 几乎处处成立. 那么对每个 $\varepsilon>0$, 以及 $0<\alpha\leqslant 1$, 有

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\left\{\sup_{g\in\mathcal{G}}\frac{\mathrm{E}(g)-\frac{1}{n}\sum_{i=1}^n g(z_i)}{\sqrt{\mathrm{E}(g)+\varepsilon}}\geqslant 4\alpha\sqrt{\varepsilon}\right\}\leqslant\mathcal{N}(\mathcal{G},\alpha\varepsilon)\exp\left\{-\frac{\alpha^2 n\varepsilon}{2c_0+\frac{2}{3}B}\right\}.$$

对如下集合应用引理 4

$$\mathcal{G}=\{g:g(z)=(f(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2,f\in\mathcal{F}_d\},$$

其中 $\mathcal{F}_d$ 如第 2 节所定义, 我们得到下面的定理 3.

定理 3 对于所有的 $\varepsilon>0$, 则有

$$\begin{aligned}\mathrm{Prob}_{\boldsymbol{z}\in Z^n}&\left\{\sup_{f\in\mathcal{F}_d}\frac{\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f)-\mathcal{E}_{\boldsymbol{z}}(m))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(m)+\varepsilon}}\leqslant\sqrt{\varepsilon}\right\}\\&\geqslant 1-\exp\left\{(2d)^s\log\left(\frac{4M^2}{\varepsilon}\right)-\frac{n\varepsilon}{32M^2}\right\}.\end{aligned}$$

证明 考虑函数集合 $\mathcal{G}$. 对于函数 $g\in\mathcal{G}$, 即 $g(z)=(f(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2$ 其中 $f\in\mathcal{F}_d$, 并满足 $\mathrm{E}(g)=\mathcal{E}(f)-\mathcal{E}(m)\geqslant 0$. 而

$$g(z)=(f(\boldsymbol{x})-y)^2-(m(\boldsymbol{x})-y)^2=(f(\boldsymbol{x})-m(\boldsymbol{z}))(f(\boldsymbol{x})+m(\boldsymbol{x})-2y),$$

对于任意的 $\boldsymbol{x}\in[-1,1]^s$, 满足 $|f(\boldsymbol{x})|\leqslant M/4$ 以及 $|m(\boldsymbol{x})|\leqslant M/4$, 我们得到 $|g(z)|\leqslant M/2\times M=M^2/2$. 于是, $|g(z)-\mathrm{E}(g)|\leqslant M^2$ 成立.

取 $c_0=M^2/2$, $B=M^2$. 对 $\alpha=\frac{1}{4}$, 由引理 4 可知, 对于每个 $\varepsilon>0$, 至少依概率

$$1-\mathcal{N}\left(\mathcal{G},\frac{\varepsilon}{4}\right)\exp\left\{-\frac{n\varepsilon}{32M^2}\right\},$$

成立

$$\sup_{f\in\mathcal{G}}\frac{\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f)-\mathcal{E}_{\boldsymbol{z}}(m))}{\sqrt{\mathcal{E}(f)-\mathcal{E}(m)+\varepsilon}}\leqslant\sqrt{\varepsilon}.$$

由函数 $g(z)$ 的表达式知,

$$\begin{aligned}|g_1(z)-g_2(z)|&\leqslant|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})||2y-f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|\\&\leqslant 4M|f_1(\boldsymbol{x})-f_2(\boldsymbol{x})|.\end{aligned}$$

因此 $\|g_1-g_2\|_\infty\leqslant M\|f_1-f_2\|_\infty$. 利用 (7) 式可以得到

$$\log\mathcal{N}\left(\mathcal{G},\frac{\varepsilon}{4}\right)\leqslant\log\mathcal{N}\left(\mathcal{F}_d,\frac{\varepsilon}{M}\right)\leqslant(2d)^s\log\left(\frac{4M^2}{\varepsilon}\right).$$

定理 3 获证.

下面利用定理 3, 给出定理 1 的证明.

定理 1 的证明 可将误差分解为

$$\begin{aligned}&\int_X |f_{\boldsymbol{z}}(\boldsymbol{x}) - m(\boldsymbol{x})|^2 \nu(\mathrm{d}\boldsymbol{x})\\ &\quad\leqslant \{(\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m)) - (\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(m))\} + (\mathcal{E}(Q_d(m)) - \mathcal{E}(m))\\ &\qquad + \{(\mathcal{E}_{\boldsymbol{z}}(Q_d(m)) - \mathcal{E}_{\boldsymbol{z}}(m)) - (\mathcal{E}(Q_d(m)) - \mathcal{E}(m))\}\\ &\quad= T_1 + T_2 + T_3.\end{aligned}$$

首先, 估计 (8) 式中的 T_1. 根据定理 3, 我们知道, 对于任意 $f \in \mathcal{F}_d$, 下面的不等式至少依概率

$$1 - \exp\left\{(2d)^s \log\left(\frac{4M^2}{t}\right) - \frac{mt}{32M^2}\right\}$$

成立

$$\mathcal{E}(f) - \mathcal{E}(m) - (\mathcal{E}_{\boldsymbol{z}}(f) - \mathcal{E}_{\boldsymbol{z}}(m)) \leqslant \sqrt{t}\sqrt{\mathcal{E}(f) - \mathcal{E}(m) + t}.$$

利用基本不等式

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

得到

$$\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m) - (\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}}) - \mathcal{E}_{\boldsymbol{z}}(m)) \leqslant t + \frac{1}{2}(\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(m))$$

至少以概率

$$1 - \exp\left\{(2d)^s \log\left(\frac{4M^2}{t}\right) - \frac{nt}{32M^2}\right\}$$

成立.

下面我们需要去求解如下方程的正解 ε_0

$$h(\varepsilon) = (2d)^s \log\left(\frac{4M^2}{\varepsilon}\right) - \frac{n\varepsilon}{32M^2} = \log\frac{\delta}{2}.$$

函数 $h : \mathcal{R}_+ \to \mathcal{R}$ 为严格递减的. 若 $h(\varepsilon^*) \leqslant \log\frac{\delta}{2}$, 则 $\varepsilon_0 \leqslant \varepsilon^*$.

当 $\varepsilon \geqslant \frac{1}{n}$, 就有

$$h(\varepsilon) \leqslant (2d)^s \log(4M^2 n) - \frac{n\varepsilon}{32M^2}.$$

如果我们取满足 $\varepsilon^* \geqslant \frac{1}{n}$ 并满足如下不等式的 ε^*

$$(2d)^s \log(4M^2 n) - \frac{n\varepsilon}{32M^2} \leqslant \log\frac{\delta}{2},$$

则有 $h(\varepsilon^*) \leqslant \log\frac{\delta}{2}$.

对充分大的 $d > 0$, 我们有

$$\varepsilon^* \geqslant \frac{32M^2}{n}\log\frac{2}{\delta} + \frac{32M^2(2d)^s \log(4M^2 n)}{n} \geqslant \frac{1}{n}.$$

从而就有

$$\varepsilon_0 \leqslant \frac{32M^2}{n}\log\frac{2}{\delta} + \frac{32M^2(2d)^s \log(4M^2 n)}{n}.$$

由定理 3 可知, 不等式

$$\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)-(\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z}})-\mathcal{E}_{\boldsymbol{z}}(m))\leqslant\frac{32M^2}{n}\log\frac{2}{\delta}+\frac{32M^2(2d)^s\log(4M^2n)}{n}+\frac{1}{2}(\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)).$$

至少以 $1-\frac{\delta}{2}$ 的置信成立.

结合定理 3 以及 (8) 式, 不等式

$$\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\leqslant\frac{204M^2}{n}\log\frac{2}{\delta}+\frac{64M^2(2d)^s\log(4M^2n)}{n}+3(\mathcal{E}(Q_d(m))-\mathcal{E}(m))$$

至少以 $1-\delta$ 的置信成立.

定理 1 得证.

为证明推论 1, 我们需要估计 $\inf_{f\in\mathcal{F}_d}\int_X(f(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x})$. 从 (4) 式知,$Q_d(f,\boldsymbol{x})\in\mathcal{H}_d$. 根据命题 2, 可知 $|Q_d(f,\boldsymbol{x})|\leqslant|f(\boldsymbol{x})|+\omega_k(f,\frac{1}{d})$.

若 $f(\boldsymbol{x})$ 的 k 阶偏导数满足 Lip_C1, 那么 $\omega_k(f,\frac{1}{d})\leqslant C_k^{'}\frac{1}{d^k}$, 其中 C_k' 为依赖于 k 的常数.

由命题 2 可知: 不等式

$$\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\leqslant\frac{204M^2}{n}\log\frac{2}{\delta}+\frac{64M^2(2d)^s\log(4M^2n)}{n}+3C_k^{'}\frac{1}{d^{2k}}.$$

极小化上述不等式右边得

$$d=\left[\left(\frac{3C_k^{'}n}{2^s64M^2\log(4M^2n)}\right)^{\frac{1}{2k+s}}\right],$$

其中 $[\cdot]$ 表示取整.

当 $n\geqslant4M^2$, 成立不等式

$$\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\leqslant\frac{204M^2}{n}\log\frac{2}{\delta}+C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}$$

至少以 $1-\delta$ 的置信成立. 其中 $C_{K,s}=2(108M^22^s)^{\frac{2k}{2k+s}}+2(3C_K^{'})^{\frac{s}{2k+s}}$.

令 $t=\frac{204M^2}{n}\log\frac{2}{\delta}+C_{K,s}(\frac{\log n}{n})^{\frac{2k}{2k+s}}$, 则有

$$\delta=2\exp\left\{-\frac{t-C_{K,s}(\frac{\log n}{n})^{\frac{2k}{2k+s}}}{\frac{204M^2}{n}}\right\},$$

则上述概率不等式可以写成

$$\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\{\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant t\}\leqslant2\exp\left\{-\frac{t-C_{K,s}(\frac{\log n}{n})^{\frac{2k}{2k+s}}}{\frac{204M^2}{n}}\right\}.$$

对于 $\tau\geqslant\frac{1}{n}$, 则有

$$\begin{aligned}E\int_X(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x})&=\int_0^\infty\mathrm{Prob}_{\boldsymbol{z}\in Z^n}\{\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant t\}\mathrm{d}t\\&\leqslant\tau+\int_\tau^\infty2\exp\left\{-\frac{t-C_{K,s}(\frac{\log n}{n})^{\frac{2k}{2k+s}}}{\frac{204M^2}{n}}\right\}\mathrm{d}t.\end{aligned}$$

极小化上述不等式的右端可得

$$\tau=\frac{204M^2\log2}{n}+C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

所以

$$E\int_X (f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{z}) \leqslant \frac{408M^2\log 2}{n}+2C_{K,s}\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

推论 1 已证毕.

3 学习算法收敛速度的下界估计

本节将给出学习算法收敛速度的下界估计, 并说明推论 1 所得的界几乎是最优的. 为了给出收敛速度的下界, 下面就先介绍集合熵数的概念.

定义 3[12] 假定 F 为 Banach 空间 E 的一个有界子集, ε 为任意正数, i 为正整数, B_E 为空间 E 的单位球. 如果存在 $x_1,x_2,\ldots,x_{2^{i-1}}\in F$ 使得 $F\subset\cup_{j=1}^{2^{i-1}}(x_j+\varepsilon B_E)$, 成立, 则使得上述关系成立的最小 ε 称为集合 F 的第 i 个熵数, 记为 $e_i(F,E)$.

下面的定理 4 给出学习速度的一个下界估计.

定理 4 令 ν 为 $X=[-1,1]^s$ 上的分布, $\Theta=\{f\in L_2(\nu):\|f\|_\infty\leqslant M/4\}$. 假定存在正整数 k, 以及常数 $c_1,c_2>0$ 使得集合 Θ 的熵数 $e_i(\Theta,L_2(\nu))$ 满足 $c_1i^{-k/s}\leqslant e_i(\Theta,L_2(\nu))\leqslant c_2i^{-k/s}$, 那么对于算法 (2), 在 $X\times[-M,M]$ 上存在一分布 P 使得 $P_X=\nu$ 以及 $m\in\Theta$, 成立

$$E\int_{[-1,1]^s}(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x})\geqslant c_1\left(\frac{1}{n}\right)^{\frac{2k}{2k+s}}.$$

定理 4 的证明基于以下引理 5.

引理 5[12] 令 ν 为 X 上的分布, $\Theta\subset L_2(\nu)$, 假定存在 $r\in(0,1)$ 使得集合 Θ 的熵数满足 $e_i(\Theta,L_2(\nu))\sim i^{-1/r}$. 那么对于算法 (2) 在 $X\times[-M,M]$ 上存在一分布 P 使得 $P_X=\nu$, 并且存在 $m\in\Theta$, 常数 $\delta_0,c_3,c_4>0$ 以及满足 $\varepsilon_n\sim n^{-\frac{2}{2+r}}$ 的序列 $\{\varepsilon_n\}$, 使得对所有 $\varepsilon>0$ 及 $n\geqslant 1$ 成立

$$\mathrm{Prob}\{\boldsymbol{z}:\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant\varepsilon\}\geqslant\begin{cases}\delta_0, & \text{当 } \varepsilon<\varepsilon_n,\\ c_3e^{-c_4n\varepsilon}, & \text{当 } \varepsilon\geqslant\varepsilon_n.\end{cases}$$

下面我们利用引理 5 证明定理 4.

证明 由于集合 Θ 第 i 个熵数满足

$$c_1i^{-k/s}\leqslant e_i(\Theta,L_2(\nu))\leqslant c_2i^{-k/s}.$$

对于 $r=\frac{s}{k}$, 应用引理 5 可知, 存在满足如下等价关系的一序列 $\{\varepsilon_n\}$, $\varepsilon_n\sim n^{-\frac{2k}{2k+s}}$ 对于 $m\in\Theta$, 成立

$$\mathrm{Prob}\{\boldsymbol{z}:\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant\varepsilon\}\geqslant\begin{cases}\delta_0, & \text{当 } \varepsilon<\varepsilon_n,\\ c_1e^{-c_2n\varepsilon}, & \text{当 } \varepsilon\geqslant\varepsilon_n,\end{cases}$$

利用上述不等式导出

$$\begin{aligned}E\int_{[-1,1]^s}(f_{\boldsymbol{z}}(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x}) &= \int_0^\infty \mathrm{Prob}(\boldsymbol{z}:\mathcal{E}(f_{\boldsymbol{z}})-\mathcal{E}(m)\geqslant\varepsilon)\mathrm{d}\varepsilon\\ &\geqslant \int_0^{\varepsilon_n}\varepsilon\mathrm{d}\varepsilon+c_1\int_{\varepsilon_n}^\infty e^{-c_2n\varepsilon}\mathrm{d}\varepsilon\geqslant\delta_0\varepsilon_n+\frac{c_1}{nc_2}e^{-c_2n^{\frac{2k}{2k+s}}}\geqslant c_1\left(\frac{1}{n}\right)^{\frac{2k}{2k+s}}.\end{aligned}$$

定理 4 证毕.

4 结论

本文研究了多项式空间上具有最小二乘平方损失正则学习算法收敛速度的上、下界问题, 所获得的上、下界仅相差一对数因子, 对刻画学习算法 (2) 的收敛特性、稳定性、适定性等具有重要的意义.

我们引进多元 Jackson 型算子, 并对该算子进行深入的研究, 尤其是利用 r 阶连续模估计了该算子的逼近阶. 同时, 结合 Hilbert 空间覆盖数及其概率论中不等式得到回归学习算法收敛速度的上界估计. 特别地, 当回归函数具有一定的光滑性, 我们得到了较好的收敛阶. 为了得到学习算法收敛速度的下界估计, 又利用集合的熵数及其估计, 在一定条件下, 给出该速度的一个下界估计.

周知, 一个算法的性能往往取决于它的收敛特征、复杂性等. 本文结果不仅说明所研究的算法是收敛的, 而且给出算法本身与样本数、假设空间等之间的函数关系. 与已有的学习算法相比, 本文所得的正则化最小二乘算法泛化误差的上、下界是较优的, 文献 [13] 也得到了较优的上、下界, 但是通过积分算子的特征值作为复杂性的度量, 而一般积分算子特征值的计算往往是较为困难的. 我们利用再生核 Hilbert 空间的覆盖数作为度量估计学习速度的上界. 然后, 引进集合的熵数并以此为工具给出学习速度的下界, 而覆盖数与熵数在学习理论研究中通常作为复杂性度量而被广泛使用.

根据推论 1 及定理 4, 当回归函数 m 的 k 阶偏导数属于 $\mathrm{Lip}_C 1$, 不难得到如下不等式

$$c_1\left(\frac{1}{n}\right)^{\frac{2k}{2k+s}} \leqslant \int_X (f(\boldsymbol{x})-m(\boldsymbol{x}))^2\nu(\mathrm{d}\boldsymbol{x}) \leqslant c_2\left(\frac{\log n}{n}\right)^{\frac{2k}{2k+s}}.$$

从上述不等式, 可以看出, 当光滑回归函数满足一些光滑性假定时, 所获得收敛速度估计除一个对数因子外是最优的. 当然, 我们希望上、下界的 "阶" 是一致的, 即能够准确刻画学习算法 (2) 的本质收敛阶, 这是一个富有挑战、值得进一步研究的问题. 我们猜测, 在一定条件下的上界估计可改进为 $c_1(\frac{1}{n})^{\frac{2k}{2k+s}}$.

参考文献

1 Vapnik V. Statistical Learning Theory. New York: Wiley. 1998

2 Shawe-Taylor J, Bartlett P L, Williamson R C, et al. Structural risk minimization over data-dependent hierarchies. IEEE Trans Inform Theory, 1998, 44: 1926–1940

3 Cucker F, Smale S. On the mathematical foundations of learning. Bull Amer Math Soc, 2001, 39: 1–49

4 Cucker F, Smale S. Best choices for regularization parameters in learning theory: on the biasvariance problem. Found Comput Math, 2002, 1: 413–428

5 Wu Q, Ying Y M, Zhou D X. Learning rates of least-square regularized regression. Found Comput Math, 2006, 6: 171–192

6 Smale S, Zhou D X. Learning theory estimates via integral operators and their approximation. Constr Approx, 2007, 26: 153–172

7 Caponnetto A, DeVito E. Optimal rates for the regularized least-squares algorithm. Found Comput Math, 2007, 7: 331–368

8 Temlyakov V. Approximation in learning theory. IMI Prepr, 2005, 5: 1–42

9 Cucker F, Zhou D X. Learning Theory: An Approximation Theory Viewpoint. Cambridge: Cambridge Univ. Press. 2007

10 Tong H Z, Chen D R, Li Z P. Learning rates for regularized classifiers using multivariate polynomial kernels. J Complex, 2008, 24: 619–631

11 Zhou D X, Jetter K. Approximation with polynomial kernels and SVM classifiers. Adv Comput Math, 2006, 25: 323–344

12 Temlyakov V. Optimal estimators in learning theory. Inst Math Polish Academy Sc, 2006, 72: 341–366
13 Steinwart I, Hush D, Scovel C. Optimal rates for regularized least squares regression. In: Proceedings of the 22nd Conference on Learning Theory. Los Alamos National Laboratory Technical Report LA-UR-09-00901. 2009
14 Chen D R, Wu Q, Ying Y, et al. Support vector machine soft margin classifiers: error analysis. J Mach Learn Res, 2004, 5: 1143–1175
15 Guo Y, Bartlett P L, Shawe-Taylor J, et al. Covering numbers for support vector machines. IEEE Trans Inform Theory, 2002, 48: 239–250
16 Pontil M. A note different covering numbers in learning theory. J Complex, 2003, 19: 665–671
17 Zhou D X. The covering number in learning theory. J Complex, 2002, 18: 739–767
18 Zhou D X. Capacity of reproducing kernel spaces in learning theory. IEEE Trans Inform Theory, 2003, 49: 1734–1752
19 Smale S, Zhou D X. Estimating the approximation error in learning theory. Anal Appl, 2003, 1: 17–41
20 Xie T F, Zhou S P. Real Function Approximation. Hangzhou: University Press, 1998
21 Xu Y. Fourier series and approximation on hexagonal and triangular domains. Constr Approx, 2010, 115–138
22 Wu Q, Ying Y M, Zhou D X. Learning theory: from regression to classification. In: Topics in Multivariate Approximation and Interpolation. Amsterdam, 2004
23 Williamson R C, Smola A J, Schŏkopf 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 Zhang T. Effective dimension and genealization of kernel learning. NIPS, 2002: 454–461
25 Zhang T. Leave-one-out bounds for kernel methods. Neural Comput, 2003, 13: 1397–1437

Estimation of convergence rate for multiregression learning algorithm

XU ZongBen[1], ZHANG YongQuan[1] & CAO FeiLong[2*]

1 *Institute for information and System Sciences, Xi'an Jiaotong University, Xi'an* 710049*, China;*
2 *Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou* 310018*, China*
*E-mail: flcao@263.net

Abstract In many applications, the pre-information on regression function is always unknown. Therefore, it is necessary to learn regression function by means of some valid tools. In this paper we investigate the regression problem in learning theory, i.e., convergence rate of regression learning algorithm with least square schemes in multi-dimensional polynomial space. Our main aim is to analyze the generalization error for multi-regression problems in learning theory. By using the famous Jackson operators in approximation theory, covering number, entropy number and relative probability inequalities, we obtain the estimates of upper and lower bounds for the convergence rate of learning algorithm. In particular, it is shown that for multi-variable smooth regression function, the estimates are able to achieve almost optimal rate of convergence except for a logarithmic factor. Our results are significant for the research of convergence, stability and complexity of regression learning algorithm.

Keywords learning theory, covering number, rate of convergence, entropy number