中国科学 : 数学　2012 年　第 42 卷　第 12 期 : 1251 ∼ 1262
www.springerlink.com　math.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

正则化回归算法的快速学习率

佟宏志[①*], 陈迪荣[②], 杨凤红[③]

① 对外经济贸易大学信息学院, 北京 100029;
② 北京航空航天大学数学系, 北京 100083;
③ 中央财经大学应用数学学院, 北京 100081
E-mail: tonghz@uibe.edu.cn, drchen@buaa.edu.cn, fhyang@cufe.edu.cn

收稿日期: 2012-01-29; 接受日期: 2012-07-27; * 通信作者
国家自然科学基金 (批准号: 11171014 和 11072274)、对外经济贸易大学校级科研课题研究成果 (批准号: 08QD28) 和对外经济贸易大学教师学术创新团队资助项目

摘要　本文讨论了再生核 Hilbert 空间上一类广泛的正则化回归算法的学习率问题. 在分析算法的样本误差时, 我们利用了一种复加权的经验过程, 保证了方差与惩罚泛函同时被阈值控制, 从而避免了繁琐的迭代过程. 本文得到了比之前文献结果更为快速的学习率.

关键词　经验过程　学习率　回归算法　正则化　再生核 Hilbert 空间
MSC (2010) 主题分类　68T05, 62J02

1　引言

设 X 为 $\mathbb{R}^n$ 的一个紧子集, $Y \subset [-M, M]$, $M > 0$. ρ 为 $Z := X \times Y$ 上未知的概率测度, $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ 为服从 ρ 的一组独立同分布的取样. 给定样本 $\mathbf{z}$, 统计学习理论中的回归问题就是要找到一个函数 $f_{\mathbf{z}} : X \to \mathbb{R}$, 使得对于任何新的不可预见的输入 $x \in X$, $f_{\mathbf{z}}(x)$ 都能成为输出 y 的一个很好的近似.

正则化的学习算法体系中有两个重要的因素: 损失函数和假设空间.

损失函数 L: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ 是一可测函数, 用来度量预测值与真实值之间的偏差. 在回归问题中最为常用的损失函数是二次损失 (或最小二乘损失), 参见文献 [1–4] 等. 然而, 近年来出于应用和计算处理方面的考虑, 不同的损失函数越来越受到人们的关注 (见 [5–7]). 本文中我们将考虑如下一类广泛的损失函数:

定义 1.1　可测函数 L: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ 称为回归损失函数, 如果存在常数 $q > 0$ 和 $c_q > 0$ 使得

$$\text{(i)}\quad |L|_1 := \sup_{y \in Y, -M \leqslant t_1, t_2 \leqslant M} \left| \frac{L(y, t_1) - L(y, t_2)}{t_1 - t_2} \right| < \infty, \tag{1.1}$$

$$\text{(ii)}\quad L(y, t) \geqslant \begin{cases} L(y, M), & \text{当}\ \ t \geqslant M, \\ L(y, -M), & \text{当}\ \ t \leqslant -M, \end{cases} \tag{1.2}$$

$$\text{(iii)}\quad L(y, t) \leqslant c_q |t|^q, \quad \forall\, |t| > M. \tag{1.3}$$

英文引用格式: Tong H Z, Chen D R, Yang F H. Fast learning rates for regularized regression algorithms (in Chinese). Sci Sin Math, 2012, 42(12): 1251–1262, doi: 10.1360/012012-74

回归损失函数的典型例子包括:

(1) ε- 不敏感损失函数 (见 [8–10])

$$L_\varepsilon(y,t)=\max\{0,|y-t|-\varepsilon\}.$$

(2) L_s- 损失函数 ($1\leqslant s<\infty$)

$$L_s(y,t)=|y-t|^s.$$

特别地, 当 $s=2$ 时就是最小二乘损失.

(3) 鲁棒损失函数 (见 [11])

$$L_r(y,t)=\begin{cases}\dfrac{(y-t)^2}{2\varepsilon}+\dfrac{\varepsilon}{2}, & \text{当}\ \ |y-t|\leqslant\varepsilon,\\ |y-t|, & \text{其他.}\end{cases}$$

(4) 对数损失函数 (见 [5])

$$L_l(y,t)=-\log(4\Lambda(|y-t|)[1-\Lambda(|y-t|)]),$$

这里 $\Lambda(r)=\frac{1}{1+\mathrm{e}^{-r}}$.

容易验证上述例子均是关于 $|y-t|$ 的非降函数, 因而 (1.2) 成立. 这些例子所满足的定义 1.1 中的参数见表 1.

预测函数 f 的性能通过如下的 L- 风险 (或推广误差) 来度量,

$$\mathcal{E}(f)=\mathcal{E}^L(f):=\int_Z L(y,f(x))d\rho=\int_X\int_Y L(y,f(x))d\rho(y|x)d\rho_X,$$

这里 ρ_X 为 X 上的边缘分布, $\rho(\cdot|x)$ 是由 ρ 诱导的 x 处的条件概率测度. 我们用 f^* 表示使 L- 风险达到最小的可测函数,

$$f^*=f_L^*:=\arg\min\mathcal{E}(f),$$

它是最为理想的学习函数, 称为目标函数. 我们总假定这样的目标函数 f^* 是存在的, 否则, 对于任意 $\eta>0$, 总存在可测函数 f_η^*, 满足 $\mathcal{E}(f_\eta^*)<\inf\mathcal{E}(f)+\eta$, 此时可用 f_η^* 来替代 f^*, 只需要做些微小的改动, 本文所有的结果均可推广到这种情况. 另外, 由 (1.2) 我们可得

$$|f^*(x)|\leqslant M,\quad x\in X. \tag{1.4}$$

学习不能在真空中进行, 假设空间是学习过程所发生的函数空间. 再生核 Hilbert 空间 (RKHS) 因其具有良好的性质, 如再生性, 成为学习理论中常用的假设空间.

表 1　典型例子满足定义 1.1 中的参数

损失函数	$\lvert L\rvert_1$	q	c_q
ε- 不敏感损失函数	1	1	2
L_s- 损失函数	$s(2M)^{s-1}$	s	2^s
鲁棒损失函数	1	1	$\max\{\frac{\varepsilon}{M},2\}$
对数损失函数	1	1	2

函数 $K: X\times X\to\mathbb{R}$ 称为一个再生核, 如果它是连续、对称和半正定的, 所谓半正定是指对于任意有限个不同点 $\{x_1,x_2,\ldots,x_l\}\subset X$, 矩阵 $(K(x_i,x_j))_{i,j=1}^l$ 是半正定的. 由核 K 所决定的再生核 Hilbert 空间 (RKHS) $\mathcal{H}_K$ 就定义为 (见 [12]) 函数集 $\{K_x:=K(x,\cdot):x\in X\}$ 所张成的线性空间按照内积 $\langle\cdot,\cdot\rangle_K$:

$$\langle K_x,K_u\rangle_K=K(x,u),$$

完备化后得到的 Hilbert 空间. 而再生性是指

$$\langle K_x,g\rangle_K=g(x),\quad\forall x\in X,\quad g\in\mathcal{H}_K. \tag{1.5}$$

记 $C(X)$ 为 X 上的连续函数空间, $\|\cdot\|_\infty$ 为其上的范数. 我们知道 $\mathcal{H}_K$ 可以连续嵌入到 $C(X)$. 令

$$\kappa:=\sup_{x\in X}\sqrt{K(x,x)}<\infty,$$

由再生性 (1.5), 可得

$$\|g\|_\infty\leqslant\kappa\|g\|_K,\quad\forall g\in\mathcal{H}_K. \tag{1.6}$$

定义 1.2 本文研究的正则化回归算法定义为下述 RKHS $\mathcal{H}_K$ 上优化问题的最小者:

$$f_{\mathbf{z},\lambda}=f_{\mathbf{z},\lambda}^L:=\arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^m L(y_i,f(x_i))+\lambda\|f\|_K^2\right\}, \tag{1.7}$$

这里 $\lambda>0$ 称为正则化参数, 其通常依赖于 m: $\lambda=\lambda(m)$, 并且 $\lim_{m\to\infty}\lambda(m)=0$.

如果我们把对应于样本 $\mathbf{z}$ 的经验 L- 风险记为

$$\mathcal{E}_{\mathbf{z}}(f)=\mathcal{E}_{\mathbf{z}}^L(f):=\frac{1}{m}\sum_{i=1}^m L(y_i,f(x_i)),$$

则算法 (1.7) 就可重写为

$$f_{\mathbf{z},\lambda}:=\arg\min_{f\in\mathcal{H}_K}\{\mathcal{E}_{\mathbf{z}}(f)+\lambda\|f\|_K^2\}. \tag{1.8}$$

当样本数量 m 增大时, 估计剩余的风险值 $\mathcal{E}(f_{\mathbf{z},\lambda(m)})-\mathcal{E}(f^*)$ 是学习理论中的一个重要课题也是本文的主要目的. 一个有效的核回归方法至少应该是相容的, 即当 $m\to\infty$ 时, 以相当高的概率, $\mathcal{E}(f_{\mathbf{z},\lambda(m)})-\mathcal{E}(f^*)\to 0$. 文献 [5] 证明了一类基于核的回归算法当 $\mathcal{H}_K$ 在 $C(X)$ 中稠密时是相容的. 进一步, 文献 [4] 和 [9] 分别对最小二乘回归 (LSR) 和使用 ε- 不敏感损失函数的支持向量机回归 (SVMR) 给出了算法的收敛速度 (通常称为学习率).

我们注意到 [4] 和 [9] 均利用了文献 [13] 中提出的 “收缩技术 (shrinking technique)” 来建立关于 $\|f_{\mathbf{z},\lambda}\|_K$ 的精确界. 另外, 投影算子 (见下面的定义 2.2) 有助于得到快速的学习率[9]. 投影的方法最早出现在分类算法的误差分析中, [4] 指出, 当 $|y|\leqslant M$ a.s., 通过投影方法可得到样本误差更精确的界. 因此, [4] 提出, 但没有给出详细的说明, 通过投影方法可以改进 LSR 的学习率.

本文的主要目的是对满足定义 1.1 的一类广泛的回归损失函数建立快速的学习率, 这其中就包括了最小二乘损失和 ε- 不敏感损失. 作为结果, 我们改进了文献 [4] 和 [9]中相应算法的学习率. 同 [9] 一样, 投影算子在误差分析中起到了关键作用.

本文的其余部分安排如下: 在第 2 节中我们给出一些重要的定义和假设并给出本文的主要定理. 主要定理的证明将在第 3 节给出. 学习率的比较会在第 4 节讨论, 在第 5 节中, 我们总结了所得到的结果并给出进一步的讨论.

2 定义、假设和主要结果

本节将给出建立快速学习率所必需的一些重要定义和假设.

定义 2.1 算法体系 (1.7) 的正则 (或逼近) 误差定义为

$$D(\lambda) := \inf_{f\in\mathcal{H}_K}\{\mathcal{E}(f)-\mathcal{E}(f^*)+\lambda\|f\|_K^2\}.$$

如果我们将 (1.7) 与样本无关的极限形式记为

$$f_\lambda := \arg\min_{f\in\mathcal{H}_K}\{\mathcal{E}(f)+\lambda\|f\|_K^2\},$$

则

$$D(\lambda)=\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)+\lambda\|f_\lambda\|_K^2. \tag{2.1}$$

为了充分利用损失函数的特征, 我们引入投影算子的概念. 投影的思想已在分类算法分析中广泛的应用, 读者可参阅文献 [14–18].

定义 2.2 对于一个可测函数 $f: X\to\mathbb{R}$, 投影算子 $\pi =: \pi_M$ 定义为

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

投影算子实际上就是对函数做一修正, 以便其函数值能够落在我们所要求的范围内. 由 (1.2) 可得 $L(y,\pi(f)(x))\leqslant L(y,f(x))$, 从而可知

$$\mathcal{E}(\pi(f))\leqslant\mathcal{E}(f),\qquad \mathcal{E}_{\mathbf{z}}(\pi(f))\leqslant\mathcal{E}_{\mathbf{z}}(f). \tag{2.2}$$

这说明用 $\pi(f)$ 去逼近目标函数 f^* 要比直接用 f 的效果好. 正是基于这种考虑, 我们下面采用 $\pi(f_{\mathbf{z},\lambda})$ 而不是原函数 $f_{\mathbf{z},\lambda}$ 作为经验目标函数并分析其相应的学习率.

同以往的分析一样, 为了得到满意的学习率, 我们也需要对概率分布 ρ 和假设空间 $\mathcal{H}_K$ 作出适当的假设.

假设 2.1 我们说目标函数 f^* 能被假设空间 $\mathcal{H}_K$ 以指数 $0<\beta\leqslant 1$ 逼近, 是指如果存在某个常数 c_β 使得

$$D(\lambda)\leqslant c_\beta\lambda^\beta,\qquad \forall\,\lambda>0.$$

下一假设是关于 ρ 和 $\mathcal{H}_K$ 的方差 - 期望界条件.

假设 2.2 存在常数 $\alpha\in[0,1]$ 和 $c_\alpha\geqslant 1$, 使得对于所有 $f\in\mathcal{H}_K$,

$$\mathbb{E}\{L(y,\pi(f)(x))-L(y,f^*(x))\}^2\leqslant c_\alpha\{\mathcal{E}(\pi(f))-\mathcal{E}(f^*)\}^\alpha.$$

本文利用经验覆盖数来度量假设空间 $\mathcal{H}_K$ 的复杂度.

定义 2.3 设 $\mathcal{F}$ 为定义在 Z 上的函数集, $\mathbf{z}:=\{z_i\}_{i=1}^m\in Z^m$. $\mathcal{F}$ 上的度量 $d_{2,\mathbf{z}}$ 定义为

$$d_{2,\mathbf{z}}(f,g):=\left\{\frac{1}{m}\sum_{i=1}^m|f(z_i)-g(z_i)|^2\right\}^{1/2}.$$

对于任意的 $\tau>0$, $\mathcal{F}$ 关于 $d_{2,\mathbf{z}}$ 的覆盖数定义为

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F},\tau):=\inf\left\{l\in\mathbb{N}:\exists\{f_i\}_{i=1}^{l}\text{ 使得 }\mathcal{F}=\bigcup_{i=1}^{l}\{f\in\mathcal{F}:d_{2,\mathbf{z}}(f,f_i)\leqslant\tau\}\right\}.$$

令 $\mathcal{B}_R=\{f\in\mathcal{H}_K:\|f\|_K\leqslant R\}$. 单位球 $\mathcal{B}_1$ 的经验覆盖数定义为

$$\mathcal{N}(\tau):=\sup_{m\in\mathbb{N}}\sup_{\mathbf{x}\in X^m}\mathcal{N}_{2,\mathbf{x}}(\mathcal{B}_1,\tau).$$

假设 2.3 称 RKHS $\mathcal{H}_K$ 具有指数为 $p\in(0,2)$ 的多项式复杂度, 如果存在常数 $c_p>0$ 使得

$$\log\mathcal{N}(\tau)\leqslant c_p(1/\tau)^p,\quad\forall\,\tau>0.$$

现将本文的主要结果叙述如下, 其证明将在下节给出.

定理 2.1 如果假设 2.1–2.3 成立, 选取 $\lambda=(\frac{1}{m})^{\min\{\frac{2}{p+\beta(4-2\alpha-(1-\alpha)p)},\frac{2}{q+(2-q)\beta}\}}$, 则对于任意的 $0<\delta<1$, 至少以 $1-\delta$ 的概率成立,

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f^*)\leqslant\tilde{c}\log(3/\delta)\left(\frac{1}{m}\right)^{\min\{\frac{2\beta}{p+\beta(4-2\alpha-(1-\alpha)p)},\frac{2\beta}{q+(2-q)\beta}\}},$$

其中 $\tilde{c}$ 是一不依赖于 m 和 δ 的常数.

3 主要定理的证明

本节将给出定理 2.1 的证明. 根据 $f_{\mathbf{z},\lambda}$ 和 f_λ 的定义, 我们可以看出

$$\begin{aligned}&\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f^*)+\lambda\|f_{\mathbf{z},\lambda}\|_K^2\\&\quad\leqslant\{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}(f^*)]-[\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))-\mathcal{E}_{\mathbf{z}}(f^*)]\}\\&\qquad+\{[\mathcal{E}_{\mathbf{z}}(f_\lambda)-\mathcal{E}_{\mathbf{z}}(f^*)]-[\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)]\}+\{\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)+\lambda\|f_\lambda\|_K^2\}\\&\quad=:I_1+I_2+I_3.\end{aligned}\tag{3.1}$$

根据 (2.1), $I_3=D(\lambda)$. 为了估计 I_2, 我们需要利用下述的单侧 Bernstein 不等式 (见 [1,17]).

引理 3.1 设 ξ 为概率空间 Z 上的一个随机变量, 其数学期望为 μ, 方差为 σ^2. 如果 $|\xi-\mu|\leqslant B$ 几乎处处成立, 则对于任意 $\eta>0$, 有

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-\mu\geqslant\eta\right\}\leqslant\exp\left\{-\frac{m\eta^2}{2(\sigma^2+\frac{1}{3}B\eta)}\right\}.$$

命题 3.1 对任意的 $t>1$, 若假设 2.2 成立, 则下式至少以 $1-2\mathrm{e}^{-t}$ 的概率成立,

$$I_2\leqslant\frac{7c_q\kappa^q t}{6m}\left(\frac{D(\lambda)}{\lambda}\right)^{q/2}+\frac{8|L|_1Mt}{3m}+\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}}+D(\lambda).$$

证明 注意到

$$I_2=\{[\mathcal{E}_{\mathbf{z}}(f_\lambda)-\mathcal{E}_{\mathbf{z}}(\pi(f_\lambda))]-[\mathcal{E}(f_\lambda)-\mathcal{E}(\pi(f_\lambda))]\}$$

$$+\{[\mathcal{E}_{\mathbf{z}}(\pi(f_\lambda))-\mathcal{E}_{\mathbf{z}}(f^*)]-[\mathcal{E}(\pi(f_\lambda))-\mathcal{E}(f^*)]\}. \tag{3.2}$$

我们先来估计 (3.2) 式右端的第一项. 由 (1.6) 和 (2.1) 可知

$$\|f_\lambda\|_\infty \leqslant \kappa\|f_\lambda\|_K \leqslant \kappa\sqrt{\frac{D(\lambda)}{\lambda}}.$$

记 $\xi_1 := L(y,f_\lambda(x))-L(y,\pi(f_\lambda)(x))$, 不妨设 $|f_\lambda(x)|>M$, 因为否则的话 $\xi_1=0$. 于是 (1.2) 和 (1.3) 就蕴含了 $0\leqslant\xi_1\leqslant c_q\kappa^q(\frac{D(\lambda)}{\lambda})^{q/2}$. 因此,

$$|\xi_1-\mathbb{E}\xi_1|\leqslant c_q\kappa^q\left(\frac{D(\lambda)}{\lambda}\right)^{q/2},\quad \sigma^2(\xi_1)\leqslant c_q\kappa^q\left(\frac{D(\lambda)}{\lambda}\right)^{q/2}\mathbb{E}\xi_1.$$

对 ξ_1 应用引理 3.1, 可得以至少 $1-\mathrm{e}^{-t}$ 的概率成立,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m\xi_1(z_i)-\mathbb{E}\xi_1 &\leqslant \frac{2c_q\kappa^q t}{3m}\left(\frac{D(\lambda)}{\lambda}\right)^{q/2}+\sqrt{\frac{2t\sigma^2(\xi_1)}{m}}\\ &\leqslant \frac{2c_q\kappa^q t}{3m}\left(\frac{D(\lambda)}{\lambda}\right)^{q/2}+\frac{c_q\kappa^q t}{2m}\left(\frac{D(\lambda)}{\lambda}\right)^{q/2}+\mathbb{E}\xi_1\\ &= \frac{7c_q\kappa^q t}{6m}\left(\frac{D(\lambda)}{\lambda}\right)^{q/2}+\mathbb{E}\xi_1.\end{aligned}$$

接下来再来估计 (3.2) 式右端的第二项. 为此我们记 $\xi_2 := L(y,\pi(f_\lambda)(x))-L(y,f^*(x))$, 那么由 (1.1), (1.4) 和假设 2.2, 可知

$$|\xi_2|\leqslant 2|L|_1M,\quad \sigma^2(\xi_2)\leqslant c_\alpha(\mathbb{E}(\xi_2))^\alpha.$$

再对 ξ_2 应用引理 3.1, 我们就得到以至少 $1-\mathrm{e}^{-t}$ 的概率,

$$\frac{1}{m}\sum_{i=1}^m\xi_2(z_i)-\mathbb{E}\xi_2\leqslant\frac{8|L|_1Mt}{3m}+\sqrt{\frac{2t\sigma^2(\xi_2)}{m}}\leqslant\frac{8|L|_1Mt}{3m}+\sqrt{\frac{2c_\alpha t(\mathbb{E}\xi_2)^\alpha}{m}}.$$

注意到基本不等式

$$\frac{1}{u}+\frac{1}{v}=1,\ \ u,v>1\Rightarrow ab\leqslant\frac{1}{u}a^u+\frac{1}{v}b^v,\ \ \forall a,b>0. \tag{3.3}$$

将其应用于 $a=(\mathbb{E}\xi_2)^{\frac{\alpha}{2}}, b=(\frac{2c_\alpha t}{m})^{\frac12}$ 和 $u=\frac{2}{\alpha}$, 就有

$$\frac{1}{m}\sum_{i=1}^m\xi_2(z_i)-\mathbb{E}\xi_2\leqslant\frac{8|L|_1Mt}{3m}+\frac{\alpha}{2}\mathbb{E}\xi_2+\left(1-\frac{\alpha}{2}\right)\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}}\leqslant\frac{8|L|_1Mt}{3m}+\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}}+\mathbb{E}\xi_2.$$

将上述的估计合在一起并注意到 $\mathbb{E}\xi_1+\mathbb{E}\xi_2=\mathcal{E}(f_\lambda)-\mathcal{E}(f^*)\leqslant D(\lambda)$, 我们便证明了命题结论. □

现在我们要估计 (3.1) 中的 I_1, 这也是主要定理证明的关键步骤. 为了收缩 $f_{\mathbf{z},\lambda}$ 所在的 RKHS 球的半径, 文献 [4,9,13,19] 等采用了一种迭代技术. 本文力图避免这一繁琐的迭代过程, 为此我们考虑如下一个 $\mathcal{H}_K$ 上的复加权的经验过程[20,21],

$$\{\omega_r(f)\{[\mathcal{E}(\pi(f))-\mathcal{E}(f^*)]-[\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f^*)]\}:\ f\in\mathcal{H}_K\}, \tag{3.4}$$

这里阈值 $r>0$, $\omega_r(f)=(r+\omega(f))^{-1}$, $\omega(f)=\mathcal{E}(\pi(f))-\mathcal{E}(f^*)+\lambda\|f\|_K^2$. 不同于传统的权重函数 (见 [20,22]), $\omega(f)$ 包含了正则化项 $\lambda\|f\|_K^2$, 这使得方差和 RKHS 范数有可能同时被阈值 r 所控制.

由于 $f_{\mathbf{z},\lambda}$ 随样本 $\mathbf{z}$ 在 $\mathcal{H}_K$ 中变动, 我们需要估计经验过程 (3.4) 的上确界. 下述的集中不等式 (concentration inequality) 在 $B=1$ 时由 [23, 定理 2.3] 给出.

引理 3.2 $z_1,\dots,z_m$ 是服从 ρ 的独立同分布的随机变量. $\mathcal{F}$ 为 Z 到 $[-B,B]$ 的可数可测函数集, 若存在某一正数 σ 使得 $\mathcal{F}$ 中任意函数 g 均满足 $\mathbb{E}g=0$, $\sigma^2(g)\leqslant\sigma^2$. 记

$$\xi=\sup_{g\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}g(z_i)\right|,$$

则对于任意 $t>0$, 都有

$$\mathrm{Prob}\left\{\xi\geqslant\mathbb{E}\xi+\sqrt{\frac{2t(\sigma^2+2B\mathbb{E}\xi)}{m}}+\frac{Bt}{3m}\right\}\leqslant\mathrm{e}^{-t}.$$

引理 3.2 描述了经验过程的上确界与其期望之间的偏差.

命题 3.2 设 $r\geqslant 0$, 记

$$V_r=\sup_{f\in\mathcal{H}_K}\omega_r(f)\left|[\mathcal{E}(\pi(f))-\mathcal{E}(f^*)]-[\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f^*)]\right|.$$

如果假设 2.2 成立, 则对于任意的 $t>0$, 以至少 $1-\mathrm{e}^{-t}$ 概率成立,

$$V_r\leqslant 2\mathbb{E}V_r+\sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}}+\frac{16|L|_1Mt}{3mr}. \tag{3.5}$$

证明 为了简便起见, 我们引入如下的一些记号:

$$\begin{aligned}&g_f(z)=g_f(x,y):=L(y,\pi(f)(x))-L(y,f^*(x)),\\&h_f(z):=\omega_r(f)[\mathbb{E}g_f-g_f(z)]=\frac{1}{\mathbb{E}g_f+\lambda\|f\|_K^2+r}[\mathbb{E}g_f-g_f(z)].\end{aligned}$$

于是

$$V_r=\sup_{f\in\mathcal{H}_K}\frac{1}{m}\left|\sum_{i=1}^{m}h_f(z_i)\right|.$$

由 (1.1) 和假设 2.2, 我们可以看出

$$\begin{aligned}&\|h_f\|_\infty\leqslant\frac{2}{r}\|g_f\|_\infty\leqslant\frac{4|L|_1M}{r},\\&\sigma^2(h_f)\leqslant\frac{\mathbb{E}(g_f)^2}{(\mathbb{E}g_f+r)^2}\leqslant\frac{c_\alpha(\mathbb{E}g_f)^\alpha}{(\frac{2}{\alpha}\mathbb{E}g_f)^\alpha(\frac{2}{2-\alpha}r)^{2-\alpha}}\leqslant c_\alpha r^{\alpha-2}.\end{aligned} \tag{3.6}$$

这里在推导 (3.6) 的第二个不等式时, 我们又一次针对 $u=\frac{2}{\alpha}$, $a=(u\mathbb{E}g_f)^{1/u}$ 和 $b=(vr)^{1/v}$ 使用了基本不等式 (3.3). 于是应用引理 3.2 于 V_r, 我们便得到以至少 $1-\mathrm{e}^{-t}$ 的概率,

$$\begin{aligned}V_r&\leqslant\mathbb{E}V_r+\sqrt{\frac{2t(c_\alpha r^{\alpha-2}+\frac{8|L|_1M\mathbb{E}V_r}{r})}{m}}+\frac{4|L|_1Mt}{3mr}\\&\leqslant\mathbb{E}V_r+\sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}}+\sqrt{\frac{16t|L|_1M\mathbb{E}V_r}{mr}}+\frac{4|L|_1Mt}{3mr}\\&\leqslant 2\mathbb{E}V_r+\sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}}+\frac{16|L|_1Mt}{3mr}.\end{aligned}$$

□

下面我们希望通过限定 V_r 的数学期望来估计 V_r. 为此我们还需要做些准备工作.

定义 3.1 $\psi:\mathbb{R}^+\to\mathbb{R}^+$ 被称为是次根 (sub-root) 函数, 如果它是非降的, 并且对于任何 $r>0$, $\psi(r)/\sqrt{r}$ 是非增的.

定义 3.2 设 (Z,ρ) 为一概率空间, $\mathcal{F}$ 为定义在 Z 上的可测函数集, $\{z_i\}_{i=1}^m$ 是 m 个独立服从于 ρ 的随机变量, $\{\epsilon_i\}_{i=1}^m$ 是 m 个独立的 Rademacher 随机变量. 那么对于任意 $\eta>0$, $\mathcal{F}$ 的局部 Rademacher 平均就定义为

$$\mathrm{Rad}(\mathcal{F},m,\eta):=\mathbb{E}\sup_{f\in\mathcal{F}:\mathbb{E}f^2\leqslant\eta}\left|\frac{1}{m}\sum_{i=1}^m\epsilon_i f(z_i)\right|.$$

下面的两个引理分别来自于文献 [20] 和 [24].

引理 3.3 设 z 是一个指标集为 T 的非负随机过程, $\omega(t)$ 是定义在 T 上的一个非负非随机的函数. 若对于任何 $r>0$, 随机变量 $\sup_{t:\omega(t)\leqslant r}z(t)$ 都有有限的数学期望, 而且存在一个次根函数 ψ 满足

$$\mathbb{E}\Big[\sup_{t\in T:\omega(t)\leqslant r}z(t)\Big]\leqslant\psi(r),\quad r\geqslant r_0\geqslant 0.$$

则对于任意的 $r\geqslant r_0$,

$$\mathbb{E}\left[\sup_{t\in T}\frac{z(t)}{\omega(t)+r}\right]\leqslant 4\frac{\psi(r)}{r}.$$

引理 3.4 设 $\mathcal{F}$ 为 Z 到 $[-B,B]$ 的可测函数集. 如果对于某个 $a>0$ 和 $p\in(0,2)$,

$$\sup_{m\in\mathbb{N}}\sup_{\mathbf{z}\in Z^m}\log\mathcal{N}_{2,\mathbf{z}}(\mathcal{F},\tau)\leqslant a\tau^{-p},\quad\forall\tau>0,$$

则存在某个仅依赖于 p 的常数 c_p', 使得

$$\mathrm{Rad}(\mathcal{F},m,r)\leqslant c_p'\max\left\{r^{1/2-p/4}\left(\frac{a}{m}\right)^{1/2},B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{2/(2+p)}\right\}.$$

命题 3.3 如果假设 2.2 和 2.3 成立, 那么对于所有的 $t>0$ 和 r 满足

$$r\geqslant\max\left\{(32c_{p,\alpha})^{\frac{4}{4-2\alpha-(1-\alpha)p}}\left(\frac{1}{m\lambda^{p/2}}\right)^{\frac{2}{4-2\alpha-(1-\alpha)p}},(32c_{p,\alpha})^{\frac{2+p}{2}}\left(\frac{1}{m\lambda^{p/2}}\right),\left(\frac{32tc_\alpha}{m}\right)^{\frac{1}{2-\alpha}},\frac{64|L|_1Mt}{3m}\right\},$$

其中 $c_{p,\alpha}:=2c_p'\{c_\alpha^{1/2-p/4}(c_p|L|_1^p)^{1/2}+(2|L|_1M)^{\frac{2-p}{2+p}}(c_p|L|_1^p)^{\frac{2}{2+p}}\}$. 均有以至少 $1-\mathrm{e}^{-t}$ 的概率成立,

$$V_r\leqslant\frac{3}{4}.$$

证明 我们仍使用在命题 3.2 的证明过程中所引入的记号, 此时权重函数 $\omega(f)=\mathbb{E}g_f+\lambda\|f\|_K^2$, 因而

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}_K:\omega(f)\leqslant r}\left|\mathbb{E}g_f-\frac{1}{m}\sum_{i=1}^m g_f(z_i)\right|\right]\leqslant\mathbb{E}\left[\sup_{\|f\|_K\leqslant\sqrt{r/\lambda},\mathbb{E}g_f\leqslant r}\left|\mathbb{E}g_f-\frac{1}{m}\sum_{i=1}^m g_f(z_i)\right|\right].$$

记 $\mathcal{G}:=\{g_f:f\in\mathcal{B}_{\sqrt{\frac{r}{\lambda}}}\}$, 根据标准的对称分析 (见 [25]) 和假设 2.2,

$$\mathbb{E}\left[\sup_{f\in\mathcal{B}_{\sqrt{\frac{r}{\lambda}}},\,\mathbb{E}g_f\leqslant r}\left|\mathbb{E}g_f-\frac{1}{m}\sum_{i=1}^m g_f(z_i)\right|\right]\leqslant 2\mathrm{Rad}(\mathcal{G},m,c_\alpha r^\alpha).$$

注意到 $\|g_f\|_\infty \leqslant 2|L|_1 M$, 以及

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{G},\tau) \leqslant \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_{\sqrt{\frac{\tau}{\lambda}}}, \frac{\tau}{|L|_1}\right) = \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_1, \frac{\tau}{|L|_1}\sqrt{\frac{\lambda}{r}}\right),$$

我们便可根据假设 2.3 得出

$$\sup_{m\in\mathbb{N}}\sup_{\mathbf{z}\in Z^m} \log \mathcal{N}_{2,\mathbf{z}}(\mathcal{G},\tau) \leqslant c_p |L|_1^p \left(\frac{r}{\lambda}\right)^{p/2} \tau^{-p}.$$

因而由引理 3.4,

$$\begin{aligned}\mathrm{Rad}(\mathcal{G}, m, c_\alpha r^\alpha) &\leqslant c'_p \max\left\{c_\alpha^{1/2-p/4} r^{\alpha(1/2-p/4)} (c_p|L|_1^p)^{1/2} \left(\frac{r}{\lambda}\right)^{p/4} \left(\frac{1}{m}\right)^{1/2},\right.\\ &\qquad \left.(2|L|_1 M)^{\frac{2-p}{2+p}} (c_p|L|_1^p)^{\frac{2}{2+p}} \left(\frac{r}{\lambda}\right)^{\frac{p}{2+p}} \left(\frac{1}{m}\right)^{\frac{2}{2+p}}\right\}\\ &\leqslant \frac{c_{p,a}}{2} \max\left\{r^{\alpha/2+(1-\alpha)p/4}\left(\frac{1}{m\lambda^{p/2}}\right)^{1/2}, \left(\frac{r}{\lambda}\right)^{\frac{p}{2+p}} \left(\frac{1}{m}\right)^{\frac{2}{2+p}}\right\}.\end{aligned}$$

令 $\psi(r) := c_{p,\alpha} \max\{r^{\alpha/2+(1-\alpha)p/4}(\frac{1}{m\lambda^{p/2}})^{1/2}, (\frac{r}{\lambda})^{\frac{p}{2+p}}(\frac{1}{m})^{\frac{2}{2+p}}\}$. 容易看到 $\psi(r)$ 是次根, 而且

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}_K:\,\omega(f)\leqslant r}\left|\mathbb{E}g_f - \frac{1}{m}\sum_{i=1}^m g_f(z_i)\right|\right] \leqslant \psi(r).$$

于是由引理 3.3 便得 $\mathbb{E}V_r \leqslant 4\frac{\psi(r)}{r}$. 根据对 r 的选择, 容易验证

$$\frac{\psi(r)}{r} \leqslant \frac{1}{32}, \quad \sqrt{\frac{2tc_\alpha r^{\alpha-2}}{m}} \leqslant \frac{1}{4} \quad 和 \quad \frac{16|L|_1 Mt}{3mr} \leqslant \frac{1}{4}.$$

这样命题的结论就可根据命题 3.2 得出. □

有了上述的准备, 下面我们便可着手证明定理 2.1.

定理 2.1 的证明 取

$$r = (32c_{p,\alpha})^{\frac{4}{4-2\alpha-(1-\alpha)p}} \left(\frac{1}{m\lambda^{p/2}}\right)^{\frac{2}{4-2\alpha-(1-\alpha)p}} + (32c_{p,\alpha})^{\frac{2+p}{2}} \left(\frac{1}{m\lambda^{p/2}}\right) + \left(\frac{32tc_\alpha}{m}\right)^{\frac{1}{2-\alpha}} + \frac{64|L|_1 Mt}{3m},$$

根据命题 3.3, 以至少 $1-\mathrm{e}^{-t}$ 的概率,

$$\omega_r(f_{\mathbf{z},\lambda})\{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f^*)]\} \leqslant V_r \leqslant \frac{3}{4}.$$

故

$$\begin{aligned}I_1 &= [\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f^*)]\\ &\leqslant \frac{3}{4}\{r + \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2\}. \qquad (3.7)\end{aligned}$$

将 (3.7) 和命题 3.1 结论合并到 (3.1) 中, 我们便得到在假设 2.1–2.3 的前提下, 以至少 $1-3\mathrm{e}^{-t}$ 的概率,

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \leqslant \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) + \lambda\|f_{\mathbf{z},\lambda}\|_K^2$$

$$\leqslant 3r + \frac{14c_q\kappa^q t}{3m}\left(\frac{D(\lambda)}{\lambda}\right)^{q/2} + \frac{32|L|_1 Mt}{3m} + 4\left(\frac{2c_\alpha t}{m}\right)^{\frac{1}{2-\alpha}} + 8D(\lambda)$$
$$\leqslant \tilde{c_1} t\left\{\left(\frac{1}{m\lambda^{p/2}}\right)^{\frac{2}{4-2\alpha-(1-\alpha)p}} + \frac{1}{m\lambda^{p/2}} + \left(\frac{1}{m}\right)^{\frac{1}{2-\alpha}} + \frac{1}{m}\lambda^{\frac{(\beta-1)q}{2}} + \lambda^\beta\right\},$$

这里

$$\tilde{c_1} := 3\left\{(32c_{p,\alpha})^{\frac{4}{4-2\alpha-(1-\alpha)p}} + (32c_{p,\alpha})^{\frac{2+p}{2}} + (32c_\alpha)^{\frac{1}{2-\alpha}} + \frac{64|L|_1 M}{3}\right\} + \frac{14c_q\kappa^q + 32|L|_1 M}{3} + 8c_\beta.$$

根据对 λ 的选择, 容易验证

$$\left(\frac{1}{m}\right)^{\frac{1}{2-\alpha}} \leqslant \lambda^\beta, \quad \frac{1}{m\lambda^{p/2}} \leqslant \lambda^\beta, \quad \left(\frac{1}{m\lambda^{p/2}}\right)^{\frac{2}{4-2\alpha-(1-\alpha)p}} \leqslant \lambda^\beta, \quad \frac{1}{m}\lambda^{\frac{(\beta-1)q}{2}} \leqslant \lambda^\beta.$$

于是通过取 $\tilde{c} = 5\tilde{c_1}$ 和 $t = \log(3/\delta)$, 我们便证明了定理. □

4 与现有学习率的比较

在本节里, 我们将主要定理应用到两种最常用的回归损失函数: 最小二乘损失和 ε- 不敏感损失上, 显示出定理 2.1 中得到的学习率要快于文献中已有的相应结果.

4.1 SVMR 的学习率

对于使用 ε- 不敏感损失的 SVMR, [9] 给出了一个显式学习率:

定理 4.1 如果假设 2.1–2.3 成立. 取 $\lambda = (\frac{1}{m})^\gamma$, 则对于任意的 $\zeta > 0$ 和 $0 < \delta < 1$, 存在不依赖于 m 的常数 $\tilde{c}$, 使得以至少 $1-\delta$ 的概率,

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f^*) \leqslant \tilde{c}\left(\frac{1}{m}\right)^\theta,$$

这里

$$\gamma = \min\left\{\frac{2}{1+\beta}, \frac{2}{2\beta(2-\alpha+p)+(1-\beta)p}\right\},$$
$$\theta = \min\left\{\frac{2\beta}{1+\beta}, \frac{2\beta}{2\beta(2-\alpha+p)+(1-\beta)p} - \zeta\right\}.$$

注意到, 对于 ε- 不敏感损失 L_ε 而言, 条件 (1.3) 中的参数 $q = 1$. 容易看出

$$\frac{2\beta}{2\beta(2-\alpha+p)+(1-\beta)p} < \frac{2\beta}{p+\beta(4-2\alpha-(1-\alpha)p)}.$$

这说明本文对 SVMR 给出的学习率要快于文献 [9].

4.2 LSR 的学习率

对于使用最小二乘损失的 LSR, 我们都知道 (参见 [3,4]) 它的目标函数 f^* 就是 ρ 的回归函数, 其定义为

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

由于对于所有的 $f \in \mathcal{H}_K$,

$$\begin{aligned}|L_2(y,\pi(f)(x)) - L_2(y,f_\rho(x))|^2 &= |(y-\pi(f)(x))^2 - (y-f_\rho(x))^2| \\ &= [2y - \pi(f)(x) - f_\rho(x)]^2[\pi(f)(x) - f_\rho(x)]^2 \\ &\leqslant 16M^2[\pi(f)(x) - f_\rho(x)]^2,\end{aligned}$$

并且

$$\begin{aligned}\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho) &= \int_Z [L_2(y,\pi(f)(x)) - L_2(y,f_\rho(x))]d\rho \\ &= \int_X \int_Y [(\pi(f))^2(x) - 2y\pi(f)(x) + 2yf_\rho(x) - f_\rho^2(x)]d\rho(y|x)d\rho_X \\ &= \int_X [\pi(f)(x) - f_\rho(x)]^2 d\rho_X.\end{aligned}$$

可得 $\mathbb{E}[L_2(y,\pi(f)(x)) - L^2(y,f_\rho(x))]^2 \leqslant 16M^2(\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho))$. 这意味着假设 2.2 对于 $\alpha = 1$ 和 $c_\alpha = 16M^2$ 总成立, 同时 L_2- 损失对于 $q = 2$ 满足条件 (1.3). 因此定理 2.1 可对 LSR 给出学习率 $(\frac{1}{m})^{\min\{\beta, \frac{2\beta}{p+2\beta}\}}$. 然而, 文献 [4] 的定理 2 得出, 在假设 2.1 和 2.3 的前提下, LSR 算法的学习率为 $(\frac{1}{m})^{\frac{\beta}{1+p}}$. 虽然这一学习率是在未进行投影的前提下得到的, 但也说明了定理 2.1 所给出的学习率确实是快速的.

5 结论和进一步的讨论

本文考虑了一类广泛的正则化回归算法的学习率问题, 其中就包括了 SVMR 和 LSR 这两种最常用的回归算法. 算法的分析可作为 [9] 工作的延伸, 但是我们通过分析一个复加权的经验过程, 替代了以往繁琐的收缩迭代过程. 新的权重函数使我们可以同时控制方差和 RKHS 范数. 作为结果, 我们得到了这类学习算法的快速学习率. 特别地, SVMR 和 LSR 的学习率得到了改进.

对于 ε- 不敏感损失函数, 本文考虑的是 ε 固定的情况. 事实上, 当 ε (随样本数量) 变化时, 会得到一些有意义的结果, 比如, 其会影响到算法的稀疏性. 对这方面问题感兴趣的读者可参见 [26].

在学习理论中, LSR 是研究比较广泛、深入的一种回归算法. 在这众多的研究成果中, 还有一类结果是讨论 $\|f_{\mathbf{z},\lambda} - f_\rho\|_K$ (假设 $f_\rho \in \mathcal{H}_K$) 的学习率, 对于这类的误差分析可参见文献 [27] 等.

在本文的误差分析中投影算子起到了关键作用. 在实际应用时, 投影参数 M 的选择依赖于输出空间 Y 的有界性. 投影后的函数是一致有界的, 因而可以利用集中不等式, 但是对于一般的损失函数 L, 本文的方法是否对未投影的原函数 $f_{\mathbf{z},\lambda}$ 仍适用, 这可作为一个未来的研究课题.

致谢 非常感谢北京大学彭立中教授对本文工作的指导和帮助. 感谢审稿人对本文提出的宝贵意见和建议.

参考文献

1 Cucker F, Smale S. On the mathematical foundations of learing theory. Bull Amer Math Soc, 2001, 39: 1–49
2 Evgeniou T, Pontil M, Poggio T. Regularization networks and support vector machines. Adv Comput Math, 2000, 13: 1–50
3 Smale S, Zhou D X. Shannon sampling and fuction reconstruction from point values. Bull Amer Math Soc, 2004, 41: 279–305
4 Wu Q, Ying Y, Zhou D X. Learning rates of least-square regularized regression. Found Comput Math, 2006, 6: 171–192

5 Christmann A, Steinwart I. Consistency and robustness of kernel-based regression in convex risk minimization. Bernoulli, 2007, 13: 799–819
6 Rosasco L, De Vito E, Caponnetto A, et al. Are loss functions all the same? Neural Comput, 2004, 16: 1063–1076
7 Steinwart I. How to compare different loss functions and their risks. Constr Approx, 2007, 26: 225–287
8 Smola A, Schölkopf B. A tutorial on support vector regression. Statist Computing, 2004, 14: 199–222
9 Tong H Z, Chen D R, Peng L Z. Analysis of support vector machines regression. Found Comput Math, 2009, 9: 243–257
10 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1996
11 Huber P J. Robust Statistics. New York: John Wiley & Sons, 1981
12 Aronszajn N. Theory of reproducing kernels. Trans Amer Math Soc, 1950, 68: 337–404
13 Scovel C, Steinwart I. Fast rates for support vector machine. In: Proceedings of the 18th Annal Conference on Learning Theory. New York: Springer, 2005, 279–294
14 Bartlett P L. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. IEEE Trans Inform Theory, 1998, 44: 525–536
15 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
16 Tong H Z, Chen D R, Peng L Z. Learning rates for regularized classifiers using multivariate polynomial kernels. J Complexity, 2008, 24: 619–631
17 Wu Q, Zhou D X. SVM soft margin classifiers: Linear programming versus quadratic programming. Neural Comput, 2005, 17: 1160–1187
18 Zhou D X, Jetter K. Approximation with polynomial kernels and SVM classifiers. Adv Comput Math, 2006, 25: 323–344
19 Wu Q, Ying Y, Zhou D X. Multi-kernel regularized classifiers. J Complexity, 2007, 23: 108–134
20 Bousquet O. Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms. PhD thesis. Paris: Ecole Polytechnique, 2002
21 Steinwart I, Hush D, Scovel C. An oracle inequality for clipped regularized risk minimizers. Neural Inform Process Syst, 2007, 19: 1321–1328
22 Bartlett P L, Bousquet O, Mendelson S. Local Rademacher complexity. Ann Statist, 2005, 33: 1497–1537
23 Bousquet O. A Bennet concentration inequality and its application to suprema of empirical process. C R Acad Sci Paris, 2002, 334: 495–500
24 Blanchard G, Lugosi G, Vayatis N. On the rate of convergence of regularized boosting classifier. J Math Learn Res, 2003, 4: 861–894
25 van der Vaart A W, Wellner J A. Weak Convergence and Empirical Processes. New York: Springer, 1996
26 Xiang D H, Hu T, Zhou D X. Approximation analysis of learning algorithms for support vector regression and quantile regression. J Appl Math, 2012, Article ID 902139, 17 pages
27 Smale S, Zhou D X. Online learning with Markov sampling. Anal Appl, 2009, 7: 87–113

Fast learning rates for regularized regression algorithms

TONG HongZhi, CHEN DiRong & YANG FengHong

Abstract We consider a family of regularized regression algorithms in reproducing kernel Hilbert spaces. In the stochastic analysis of these algorithms, we avoid the prolix iteration via a reweighted empirical process which restricts the variance and the penalty functional as well. As a result, we derive some faster learning rates in comparison with ones known in the literature.

Keywords **empirical processes, learning rates, regression, regularization, reproducing kernel Hilbert spaces**

MSC(2010) **68T05, 62J02**

doi: 10.1360/012012-74