

非一致分布下的在线分位数回归算法

杨鹏伟

(武汉大学数学与统计学院, 湖北 武汉 430072)

摘要: 本文在非一致抽样分布下, 研究与高斯核有关的在线分位数算法的收敛阶. 本文引入阈值 ϵ 到 Pinball 损失函数产生算法的稀疏性, 用 Hölder 对偶空间刻画抽样分布的非一致性, 通过误差分解和迭代方法推导算法的收敛速度. 并且以中位数回归为例, 得到算法的具体收敛速度, 同时也指明本文的背景和数学方法适用于一般分位数回归.

关键词: 在线算法; 分位数回归; ϵ -pinball 损失函数; 再生核 Hilbert 空间; 非一致分布

MR(2010) 主题分类号: 62J99

中图分类号: O29

文献标识码: A

文章编号: 0255-7797(2019)05-0713-08

1 引言

在学习理论的文章中, 算法的样本 $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^T$ 通常是独立同分布的 (来自相同分布且相互独立的). 在实际问题中, 这一假设往往不成立 (具体例子在正文给出). 基于非一致抽样下的研究, 较早见于 Smale, Zhou^[1] 的工作, 主要探讨最小二乘回归下的在线学习. Steinwart 等^[2] 研究了和支持向量机有关的批量学习. Hu, Zhou^[3] 探讨了一般损失函数的非一致在线学习算法, 他们的分析可用于回归和分类等学习问题. 本文在非一致分布背景下, 研究在线分位数回归算法, 通过带有阈值 ϵ -pinball 损失函数产生稀疏性, 以中位数回归为代表, 得到算法的收敛阶.

1.1 不同分布下的样本

假设输入空间 (X, d) 是 R^n 中的紧度量空间, 输出空间 $Y \subset R$, 样本空间 $Z = X \times Y$. 非一致分布下, 每一学习时刻 $t = 1, 2, \dots$ 产生的样本点 $z_t = (x_t, y_t) \in Z$ 服从联合概率分布 $\rho^{(t)}(x, y)$, $(x, y) \in Z$, 这里序列 $\{\rho^{(t)}\}_{t=1,2,\dots}$ 不一定相同. 在 Smale, Zhou^[1] 提出的学习框架下, 假设在 X 上的边缘分布序列 $\{\rho_X^{(t)}\}_{t=1,2,\dots}$ 以多项式速度收敛于 Hölder 空间 $C^s(X)$ ($0 < s \leq 1$) 的对偶空间 $(C^s(X))^*$ ^[1,3], Hölder 空间 $C^s(X)$ 是由定义在 X 上所有连续函数组成的空间, 其范数

$$\|f\|_{C^s(X)} = \|f\|_{C(X)} + |f|_{C^s(X)}, \quad |f|_{C^s(X)} := \sup_{\substack{x \neq x', \\ x, x' \in X}} \frac{|f(x) - f(x')|}{(d(x, x'))^s}, \quad 0 \leq s \leq 1.$$

定义 1.1 称概率分布序列 $\{\rho_X^{(t)}\}_{t=1,2,\dots}$ 以多项式速度收敛于 $(C^s(X))^*$ 中的概率分布 ρ_X . 如果存在 $C > 0, b > 0$ 使得

$$\left\| \rho_X^{(t)} - \rho_X \right\|_{(C^s(X))^*} \leq Ct^{-b}, \quad t \in \mathbb{N}. \quad (1.1)$$

*收稿日期: 2019-02-25 接收日期: 2019-05-05

基金项目: 国家自然科学基金资助 (11671307).

作者简介: 杨鹏伟 (1992-), 男, 河南安阳, 硕士, 主要研究方向: 统计学习基础.

由对偶空间 $(C^s(X))^*$ 的定义, 条件 (1.1) 等价于

$$\left| \int_X f(x) d\rho_X^{(t)} - \int_X f(x) d\rho_X \right| \leq Ct^{-b} \|f\|_{C^s(X)}, \forall f \in C^s(X), t \in \mathbb{N}, \quad (1.2)$$

指数 b 衡量不同分布的差异程度, 当 $b = \infty$ 时, 条件 (1.1) 是独立同分布.

对于满足衰减条件 (1.1) 的实际情况已在文献 [1, 3] 有详细讨论. 例如, 真实的抽样分布 ρ_X 被噪声干扰并且随着时间 t 增加而噪声水平下降, 那么在不同时刻对应的抽样概率分布 $\{\rho_X^{(t)}\}$ 收敛于 ρ_X . 另外, 通过迭代和随机密度核形成的积分算子, 反映了由动力系统诱导出不同分布的表现形式. 这些例子都说明研究满足 (1.1) 衰减的分布是具有实际应用性的.

1.2 分位数回归

对于任意的 $0 < \tau < 1$, 定义随机变量 Y 的 τ 分位数函数 $Q(\tau)$ 为

$$Q(\tau) = \inf\{y : F(y) \geq \tau\},$$

其中 $F(y)$ 是 Y 的分布函数. Koenker, Bassett^[4] 提出线性分位数回归理论, 扩展了经典的最小二乘回归. Koenker^[5] 进一步提出分位数回归可提供更多关于因变量的分布信息, 例如: 长尾性、厚尾性、多峰性. 分位数回归描述因变量的条件分布, 不仅仅分析因变量的条件期望, 并且对不同 τ 刻画了概率分布的具体性质.

在学习理论框架下, ρ 是 Z 上的一个 Borel 概率测度. 给定 $x \in X$ 时, $\rho_x(y), y \in Y$ 是 ρ 的条件概率分布. 分位数回归的目标函数 $f_{\rho, \tau}(x)$ 在 x 的值定义为: $\rho_x(\cdot)$ 的 τ -分位数是 v , 即存在 $v \in Y$ 满足

$$\rho_x(\{y \in Y : y \leq v\}) \geq \tau \text{ 且 } \rho_x(\{y \in Y : y \geq v\}) \geq 1 - \tau. \quad (1.3)$$

当 $\tau = \frac{1}{2}$ 时, 分位数回归就是中位数回归, 中位数回归在分位数回归中占有相当重要的地位. 我们知道, 当 $\rho_x(\cdot)$ 关于中位数对称时, 中位数回归即为条件概率的均值, 对应于经典最小二乘法. 一般分位数回归, 对应的损失函数是 pinball 损失 $V = V^\tau : \mathbb{R} \rightarrow \mathbb{R}_+$ (见文献 [6]),

$$V^\tau(u) = \begin{cases} (1 - \tau)u, & u > 0, \\ -\tau u, & u \leq 0. \end{cases} \quad (1.4)$$

为了对分位数回归算法产生稀疏性, 引入阈值 $\epsilon > 0$ 生成的 ϵ -pinball 损失 V_ϵ^τ 定义为

$$V_\epsilon^\tau(u) = \begin{cases} (1 - \tau)(u - \epsilon), & u > \epsilon, \\ -\tau(u + \epsilon), & u \leq -\epsilon, \\ 0, & \text{其它}. \end{cases} \quad (1.5)$$

当 $\tau = \frac{1}{2}$, V_ϵ^τ 就是 ϵ -不敏感损失函数. 当估计值与真值的偏差小于 ϵ 时, 损失为零. 当 ϵ 比较大时, 导致结果的正确率会略有下降, 但是数据的训练、测试和参数选择速度均会提高, 尤其在大规模数据集上, 这种优势会更加明显. ϵ -pinball 损失函数 V_ϵ^τ 在支持向量机算法中应用广泛, 因此将 V_ϵ^τ 和在线学习算法结合研究并给出收敛阶, 具有实际应用价值.

1.3 在线分位数回归算法

再生核 Hilbert 空间 (RKHS) 是由 Mercer 核 $K: X \times X \rightarrow \mathbb{R}$ 诱导出来的, K 是一个连续对称函数, 并且对于任意有限点集 $\{x_1, x_2, \dots, x_l\} \subset X$ 生成的矩阵 $(K(x_i, x_j))_{i,j=1}^l$ 是半正定的. 由函数集 $\{K_x = K(x, \cdot) : x \in X\}$ 张成的完备线性闭包空间 RKHS 记为 $\mathcal{H}_K$ (Aronszajn, 1950^[7]), 记内积为 $\langle K_x, K_y \rangle_K = K(x, y)$. 在线算法也称随机梯度下降算法 (SGD) 与批次算法中样本一次性传给机器不同, 在线算法的训练样本是按时间 t 顺序依次传送给机器, 对算法不断修正, 提高学习效率. 在线算法由于低复杂度, 在流式数据和大规模计算中有广泛的应用. 具体可参考文献 [8].

定义 1.2 与 RKHS 有关的在线分位数回归算法定义为

$$\begin{aligned} f_1 &= 0, \\ f_{t+1} &= \begin{cases} (1 - \lambda_t \eta_t) f_t - (1 - \tau) \eta_t K_{x_t}, & f_t(x_t) - y_t > \epsilon, \\ (1 - \lambda_t \eta_t) f_t + \tau \eta_t K_{x_t}, & f_t(x_t) - y_t \leq -\epsilon, \\ (1 - \lambda_t \eta_t) f_t, & -\epsilon < f_t(x_t) - y_t \leq \epsilon, \end{cases} \end{aligned} \quad (1.6)$$

其中 t 是学习时间, $\lambda_t > 0$ 称为正则参数, $\eta_t > 0$ 是步长.

在线算法 (1.6) 中, 正则参数 λ_t 随着学习时间 t 变化并且 $\lambda_{t+1} \leq \lambda_t, \forall t \in \mathbb{N}$. 当 $\lambda_t \equiv \lambda_1$ 时, λ_t 不会随着步数 t 变化而变化, 此时称 (1.6) 为不完全在线算法.

2 非一致分布下在线分位数回归的误差

高斯函数在统计学领域, 用于表述正态分布; 在信号处理领域, 用于定义高斯滤波器; 在图像处理领域, 二维高斯核函数常用于高斯模糊; 在数学领域, 主要是用于解决热力学方程和扩散方程. 高斯核函数作为最常用的径向基函数, 在支持向量机等算法中的应用可以将数据映射到高维甚至无穷维. 由于高斯核具有良好的性质和应用广泛, 因此将高斯核应用于以下与中位数回归有关的在线算法 (1.6) 中.

对于 $f: X \rightarrow \mathbb{R}$ 的泛化误差定义为

$$\mathcal{E}^\epsilon(f) = \int_Z V_\epsilon^\tau(y - f(x)) d\rho = \int_X \int_Y V_\epsilon^\tau(y - f(x)) d\rho_x(y) d\rho_X,$$

其中 $f_\rho^{V_\epsilon^\tau}$ 是 $\mathcal{E}^\epsilon(f)$ 在全体可测函数中的极小化值. 特别的, 当 $\epsilon = 0$, 记 $\mathcal{E}(f) := \mathcal{E}^0(f)$, $f_\rho^{V_\tau} := f_\rho^{V_0^\tau}$. 本文中, $f_\rho^{V_\tau}$ 是算法 (1.6) 的目标函数. 以下定理提供了算法 (1.6) 的学习收敛率, 其证明将在 2.2 节给出.

定理 2.1 假设以下假设均成立

1. RKHS $\mathcal{H}_K$ 由高斯函数 $K(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$, $\forall x, x' \in X$, $\sigma > 0$ 产生;
2. 分位数参数 $\tau = \frac{1}{2}$;
3. 正则条件 $f_\rho^{V_\tau} = L_K^r(g_\rho)$, 其中 L_K^r 是紧算子, $g_\rho \in L_{\rho_X}^2(X)$, L_K 是积分算子 $L_{\rho_X}^2$: $L_K f(x) = \int_X K(x, v) f(v) d\rho_X(v)$, $x \in X$, 成立且 $r > \frac{1}{2}$;
4. 关于 $\{\rho_X^{(t)}\}$ 的多项式收敛条件 (1.1) 成立;
5. $\forall x \in X$, 条件分布 $\rho_x(\cdot)$ 是区间 $[f_\rho(x) - 1, f_\rho(x) + 1]$ 上的一致分布.

如果 $\lambda_1 < \frac{(\|g_\rho\|_{L^2_{\rho_X}})^{\frac{2}{1-2r}}}{2}$, $\eta_1 > 0$, $0 \leq \epsilon \leq \lambda_T^{\frac{2}{r}}$, 则有
 当 $1 < r \leq \frac{3}{2}$, $\lambda_T = \lambda_1 T^{-\frac{2}{6r+1}}$, $\eta_T = \eta_1 T^{-\frac{4r}{6r+1}}$ 时,

$$\mathbb{E}_{z_1, z_2, \dots, z_T} (\|f_{T+1} - f_\rho^{V^\tau}\|_K) \leq \tilde{C}_1 T^{-\min\{\frac{2r-1}{6r+1}, \frac{b}{2} - \frac{2}{6r+1}\}}. \quad (2.1)$$

当 $\frac{1}{2} < r \leq 1$, $\lambda_T = \lambda_1 T^{-\frac{1}{3r+2}}$, $\eta_T = \eta_1 T^{-\frac{2r+1}{3r+2}}$ 时,

$$\mathbb{E}_{z_1, z_2, \dots, z_T} (\|f_{T+1} - f_\rho^{V^\tau}\|_{L^2_{\rho_X}}) \leq \tilde{C}_2 T^{-\min\{\frac{r}{3r+2}, \frac{b}{2} - \frac{1}{3r+2}\}}, \quad (2.2)$$

其中 $\tilde{C}_1$ 和 $\tilde{C}_2$ 是与 T 无关的常数, 将在证明过程中给出.

注 1 该定理反应了中位数回归的在线算法收敛速度, (2.1) 式说明了在线算法 (1.6) 在 $\mathcal{H}_K$ 的收敛性, 称为强收敛; 而结果 (2.2) 称为算法的平均误差. 由范数关系 $\|\cdot\|_{L^2_{\rho_X}} \leq \|\cdot\|_K$ 知道本定理结果 (2.1) 和 (2.2) 是合理的. 虽然本定理是考虑 $\tau = \frac{1}{2}$, 但是从后面的证明过程知道本定理结果可以推广到任意 $0 < \tau < 1$, 只需要修改常数 $\tilde{C}_1$ 和 $\tilde{C}_2$.

注 2 在线算法 (1.6) 的误差估计分为抽样误差和逼近误差两部分, 其中抽样误差是本文的主要贡献, 将在定理 2.3 中给出; 逼近误差由 $\mathcal{H}_K$ 空间的复杂度和数据的分布 ρ 决定. 假设 5 采用一致分布估计逼近误差. 事实上本定理适用于其他更一般的条件分布, 为了简化本定理, 不再进一步讨论. 具体可参考文献 [2, 3].

注 3 当 $b = \infty$, (2.1) 和 (2.2) 式反映了数据抽样 $\mathbf{z}$ 是独立同分布的情况. 另外, 我们看到 $\epsilon = 0$ 时, 本定理推导出无稀疏性的在线分位数回归算法的收敛阶, 可以通过调节 ϵ 的值控制算法的稀疏性同时不影响算法的收敛率.

2.1 抽样误差的估计

定义 2.2 对于 $\lambda > 0$, 正则化函数 $f_\lambda^{V^\tau} \in \mathcal{H}_K$ 定义为

$$f_\lambda^{V^\tau} := \arg \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}^\epsilon(f) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \quad (2.3)$$

逼近误差 $\mathcal{D}(\lambda)$ 定义为

$$\mathcal{D}(\lambda) := \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho^{V^\tau}) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \quad (2.4)$$

以下定理给出一般分位数回归算法的抽样误差 $f_{T+1} - f_{\lambda_T}^{V^\tau}$.

定理 2.3 如果以下假设成立

1. 核函数 K 与定理 2.1 相同;
2. $\{\rho_X^{(t)}\}_{t=1,2,\dots}$ 以多项式速度收敛到 $(C^s(X))^*$ 中的 ρ_X , 即 (1.1) 成立;
3. 分布 $\{\rho_x : x \in X\}$ 在 $(C^s(X))^*$ 中是 Lipschitz s 的, 即存在一个常数 $C_\rho \geq 0$, 使得

$$\|\rho_x - \rho_u\|_{(C^s(Y))^*} \leq C_\rho (d(x, u))^s, \quad \forall x, u \in X. \quad (2.5)$$

4. 逼近误差满足多项式衰减如下

$$\mathcal{D}(\lambda) \leq \mathcal{D}_0 \lambda^\beta, \quad 0 < \beta \leq 1, \mathcal{D}_0 > 0. \quad (2.6)$$

令

$$\begin{aligned} \lambda_t &= \lambda_1 t^{-\gamma}, \quad \eta_t = \eta_1 t^{-\alpha}, \quad 0 < \lambda_1, \eta_1 < 1, \\ 0 < \gamma &< \frac{2}{5-\beta}, \quad \gamma < \alpha < 1 - \frac{\gamma(3-\beta)}{2}, \quad \epsilon \leq \lambda^\beta, \end{aligned}$$

则有

$$\mathbb{E}_{z_1, z_2, \dots, z_T} (\|f_{T+1} - f_{\lambda_T}^{V_\epsilon^\tau}\|_K^2) \leq C_{K, \rho, b, \beta} T^{-\theta},$$

其中

$$\theta := \min\{2 - \gamma(3 - \beta) - 2\alpha, \alpha - \gamma, b - 2\gamma\},$$

$C_{K, \rho, b, \beta}$ 是独立于 T 的一个常数且在证明中给出.

证 本定理的证明借助了 Hu, Zhou^[3] 的证明思想. 以下给出证明的主要步骤, 更详细的过程参见 Hu, Zhou^[3] 的定理 26. 证明分为三个步骤.

第一步 存在常数 $\kappa_{2s} > 0$,

$$|K(x, x) - 2K(x, u) + K(u, u)| \leq \kappa_{2s}^2 (d(x, u))^{2s}, \quad \forall x, u \in X.$$

由 $K(x, x) = K(u, u) = 1$, 得

$$|K(x, x) - 2K(x, u) + K(u, u)| = 2 - 2e^{-\frac{\|x-u\|^2}{2\sigma^2}} \leq \frac{\|x-u\|^2}{\sigma^2}. \quad (2.7)$$

又 $(d(x, u))^{2s} = \|x - u\|^{2s}$, 可令 $s = 1$, 即存在 $\kappa_{2s} \geq \frac{1}{\sigma} > 0$, 使得 (2.7) 式成立.

第二步 由 Ying, Zhou^[8] 引理 3 知

$$\mathbb{E}_{z_t} (\|f_{t+1} - f_{\lambda_t}^{V_\epsilon^\tau}\|_K^2) \leq (1 - \eta_t \lambda_t) \|f_t - f_{\lambda_t}^{V_\epsilon^\tau}\|_K^2 + 2\eta_t \Delta_t + \eta_t^2 \mathbb{E}_{z_t} \|\partial V_\epsilon^\tau(y_t, f_t(x_t)) K_{x_t} + \lambda_t f_t\|_K^2, \quad (2.8)$$

其中

$$\Delta_t = \int_Z \{V_\epsilon^\tau(y, f_{\lambda_t}^{V_\epsilon^\tau}(x)) - V_\epsilon^\tau(y, f_t(x))\} d[\rho^{(t)} - \rho].$$

因此需要估计 Δ_t 的上界 (参考文献 [3]). 不难证明存在常数 $C_{V_\epsilon^\tau, \gamma, \alpha} > 0$, 使得 $\|f_t\|_\infty \leq \frac{C_{V_\epsilon^\tau, \gamma, \alpha}}{\lambda_t}$, $t = 1, 2, \dots, T+1$. 根据 V_ϵ^τ 关于第二变量的一阶 Taylor 展开, 可知

$$\begin{aligned} \Delta_t &\leq \left\{ (1 + \kappa_{2s}) \left(\sqrt{\frac{2\|V_\epsilon^\tau\|_\infty}{\lambda_t}} + \frac{C_{V_\epsilon^\tau, \gamma, \alpha}}{\lambda_t} \right) N(\lambda_t) + 2C_\rho \tilde{N}(\lambda_t) \right\} \|\rho_X^{(t)} - \rho_X\|_{(C^s(X))^*} \\ &\leq B_t^* := Ct^{-b} \left\{ (1 + \kappa_{2s}) \left(\sqrt{\frac{2\|V_\epsilon^\tau\|_\infty}{\lambda_t}} + \frac{C_{V_\epsilon^\tau, \gamma, \alpha}}{\lambda_t} \right) N(\lambda_t) + 2C_\rho \tilde{N}(\lambda_t) \right\}, \end{aligned} \quad (2.9)$$

这里

$$N(\lambda) = \sup \left\{ |\partial V_\epsilon^\tau(y, f)| : y \in Y, |f| \leq \max \left\{ C_{V_\epsilon^\tau, \gamma, \alpha} / \lambda, \kappa \sqrt{2 \|V_\epsilon^\tau\|_\infty / \lambda} \right\} \right\},$$

$$\tilde{N}(\lambda) = \sup \left\{ \|V_\epsilon^\tau(y, f)\|_{C^s(Y)} : y \in Y, |f| \leq \max \left\{ C_{V_\epsilon^\tau, \gamma, \alpha} / \lambda, \kappa \sqrt{2 \|V_\epsilon^\tau\|_\infty / \lambda} \right\} \right\}.$$

将估计 (2.9) 式代入 (2.8) 式得

$$\mathbb{E}(\|f_{t+1} - f_{\lambda_t}^{V_\epsilon^\tau}\|_K^2) \leq (1 - \eta_t \lambda_t) \mathbb{E}_{z_1, z_2, \dots, z_{t-1}}(\|f_t - f_{\lambda_t}^{V_\epsilon^\tau}\|_K^2) + \eta_t^2 (N(\lambda_t) + C_{V_\epsilon^\tau, \gamma, \alpha})^2 + 2\eta_t B_t^*. \quad (2.10)$$

第三步 根据 $\|\partial V\|_\infty < \infty$, 知必然存在常数 C' , 使得 $N(\lambda) \leq C', \tilde{N}(\lambda) \leq C'\lambda$. 同时, 注意到当 $\epsilon \leq \lambda^\beta$, 可以引用文献 [9] 中引理 3(b), 得到

$$\mathcal{D}_\epsilon(\lambda) := \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho^{V_\epsilon^\tau}) + \frac{\lambda}{2} \|f\|_K^2 \right\} \leq \mathcal{D}(\lambda) + 2\epsilon \leq (\mathcal{D}_0 + 2)\lambda^\beta.$$

所以用文献 [3] 定理 20 的结果, 得到

$$\|f_{\lambda_t}^{V_\epsilon^\tau} - f_{\lambda_{t-1}}^{V_\epsilon^\tau}\|_K \leq 2\sqrt{\mathcal{D}_0 + 2\lambda_1^{\frac{\beta-1}{2}}} (t-1)^{\frac{\gamma(1-\beta)}{2}-1} \leq 4\sqrt{\mathcal{D}_0 + 2\lambda_1^{\frac{\beta-1}{2}}} t^{\frac{\gamma(1-\beta)}{2}-1}.$$

将上述估计插入 (2.10) 式, 并按 $t = T, \dots, 1$ 顺序进行迭代, 得到本定理结论.

2.2 定理 2.1 的证明

证 将误差 $f_{T+1} - f_\rho^{V_\epsilon^\tau}$ 分解为三项: $f_{T+1} - f_{\lambda_T}^{V_\epsilon^\tau}$, $f_{\lambda_T}^{V_\epsilon^\tau} - f_{\lambda_T}^{V_\tau}$ 和 $f_{\lambda_T}^{V_\tau} - f_\rho^{V_\tau}$.

首先, 根据条件 5 和在文献 [10] 例 6 知道, 在正则条件 $f_\rho^{V_\tau} = L_K^\tau(g_\rho)$ 下

$$\|f_{\lambda_T}^{V_\tau} - f_\rho^{V_\tau}\|_K \leq \left(\frac{\lambda_T}{2}\right)^{r-\frac{1}{2}} \|g_\rho\|_{L_{\rho_X}^2}, \quad \frac{1}{2} < r \leq \frac{3}{2},$$

$$\|f_{\lambda_T}^{V_\tau} - f_\rho^{V_\tau}\|_{L_{\rho_X}^2} \leq \left(\frac{\lambda_T}{2}\right)^r \|g_\rho\|_{L_{\rho_X}^2}, \quad 0 < r \leq 1.$$

接下来, 关于第二项 $f_{\lambda_T}^{V_\epsilon^\tau} - f_{\lambda_T}^{V_\tau}$ 的估计, 由 Hu, Xiang 等^[11] 性质 1 知

$$\|f_{\lambda_T}^{V_\epsilon^\tau} - f_{\lambda_T}^{V_\tau}\|_K \leq C_\rho \epsilon^s \lambda_T^{-1} = C_\rho \lambda_1^{-1} \epsilon^s T^\gamma.$$

关于第一项 $f_{T+1} - f_{\lambda_T}^{V_\epsilon^\tau}$ 的估计需要借助定理 2.3. 逐条验证定理 2.3 的条件 1-4, 显然在定理 2.1 的条件下, 定理 2.3 的条件 1-2 成立.

注意到 $\|f_\rho^{V_\tau}\|_K \leq \|g_\rho\|_{L_{\rho_X}^2}$ 和 $\|f_\rho^{V_\tau}\|_{C^s(X)} \leq (1 + \kappa_{2s})\|f_\rho\|_K$, ρ_x, ρ_u 均为均匀分布, 则 $\forall x, u \in X, g \in C^s(Y)$, 有

$$\left| \int_Y g(y) d(\rho_x - \rho_u) \right| = \frac{1}{2} \left| \int_{f_\rho(x)-1}^{f_\rho(x)+1} g(y) dy - \int_{f_\rho(u)-1}^{f_\rho(u)+1} g(y) dy \right|$$

$$\leq \|g\|_{C(Y)} |f_\rho(x) - f_\rho(u)|$$

$$\leq \|g\|_{C(Y)} (1 + \kappa_2) \|g_\rho\|_{L_{\rho_X}^2} (d(x, u)).$$

Lipschitz s 连续条件 (2.5), 其中 $C_\rho = (1 + \kappa_{2s})\|g_\rho\|_{L^2_{\rho_X}}$. 即定理 2.3 的条件 3 满足.

关于定理 2.3 的条件 4, 因为正则指数 $r \geq \frac{1}{2}$, $f_\rho^{V^\tau} \in \mathcal{H}_K$, 所以 $\mathcal{D}(\lambda) \leq \lambda\|f_\rho^{V^\tau}\|_K^2$, 即 (2.6) 式成立且 $\beta = 1$, $\mathcal{D}_0 = \|f_\rho^{V^\tau}\|_K^2$. 综上, 可得

$$\mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_{\lambda_T}^{V^\tau}\|_K^2) \leq C_{K, \rho, b, \beta} T^{-\theta^*},$$

这里 $\theta^* := \min\{2 - 2\gamma - 2\alpha, \alpha - \gamma, b - 2\gamma\}$.

根据以上推导得到

$$\begin{aligned} & \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_\rho^{V^\tau}\|_K) \\ & \leq \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_{\lambda_T}^{V^\tau}\|_K) + \|f_{\lambda_T}^{V^\tau} - f_{\lambda_T}^{V^\tau}\|_K + \|f_{\lambda_T}^{V^\tau} - f_\rho^{V^\tau}\|_K \\ & \leq \sqrt{C_{K, \rho, b, \beta}} T^{-\frac{\theta^*}{2}} + C_\rho \lambda_1^{-1} \epsilon^s T^\gamma + \lambda_T^{r-\frac{1}{2}} \|g_\rho\|_{L^2_{\rho_X}}, \quad 1 < r \leq \frac{3}{2}, \end{aligned} \quad (2.11)$$

$$\begin{aligned} & \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_\rho^{V^\tau}\|_{L^2_{\rho_X}}) \\ & \leq \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_{\lambda_T}^{V^\tau}\|_{L^2_{\rho_X}}) + \|f_{\lambda_T}^{V^\tau} - f_{\lambda_T}^{V^\tau}\|_{L^2_{\rho_X}} + \|f_{\lambda_T}^{V^\tau} - f_\rho^{V^\tau}\|_{L^2_{\rho_X}} \\ & \leq \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_{\lambda_T}^{V^\tau}\|_K) + \|f_{\lambda_T}^{V^\tau} - f_{\lambda_T}^{V^\tau}\|_K + \|f_{\lambda_T}^{V^\tau} - f_\rho^{V^\tau}\|_{L^2_{\rho_X}} \\ & \leq \sqrt{C_{K, \rho, b, \beta}} T^{-\frac{\theta^*}{2}} + C_\rho \lambda_1^{-1} \epsilon^s T^\gamma + \lambda_T^r \|g_\rho\|_{L^2_{\rho_X}}, \quad \frac{1}{2} < r \leq 1. \end{aligned} \quad (2.12)$$

当 $1 < r \leq \frac{3}{2}$ 时, 将 $\lambda_T = \lambda_1 T^{-\frac{2}{6r+1}}$, $\eta_T = \eta_1 T^{-\frac{4r}{6r+1}}$, $\epsilon = \lambda_T^{\frac{2}{r}}$ 代入 (2.11) 式得

$$\begin{aligned} \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_\rho\|_K) & \leq \sqrt{C_{K, \rho, b, \beta}} T^{-\frac{\theta^*}{2}} + C_\rho \lambda_1 T^{-\frac{2}{6r+1}} + \lambda_1^{r-\frac{1}{2}} \|g_\rho\|_{L^2_{\rho_X}} T^{-\frac{2r-1}{6r+1}} \\ & \leq \tilde{C}_1 T^{-\min\{\frac{2r-1}{6r+1}, \frac{b}{2} - \frac{2}{6r+1}\}}, \end{aligned}$$

其中 $\tilde{C}_1 := \sqrt{C_{K, \rho, b, \beta}} + C_\rho \lambda_1 + \lambda_1^{r-\frac{1}{2}} \|g_\rho\|_{L^2_{\rho_X}}$. 结论 (2.1) 式成立.

当 $\frac{1}{2} < r \leq 1$ 时, 将 $\lambda_T = \lambda_1 T^{-\frac{1}{3r+2}}$, $\eta_T = \eta_1 T^{-\frac{2r+1}{3r+2}}$ 代入 (??) 式得

$$\begin{aligned} \mathbb{E}_{z_1, z_2, \dots, z_T}(\|f_{T+1} - f_\rho\|_{L^2_{\rho_X}}) & \leq \sqrt{C_{K, \rho, b, \beta}} T^{-\frac{\theta^*}{2}} + C_\rho \lambda_1 T^{-\frac{1}{3r+2}} + \lambda_1^r \|g_\rho\|_{L^2_{\rho_X}} T^{-\frac{r}{3r+2}} \\ & \leq \tilde{C}_2 T^{-\min\{\frac{r}{3r+2}, \frac{b}{2} - \frac{1}{3r+2}\}}, \end{aligned}$$

其中 $\tilde{C}_2 := \sqrt{C_{K, \rho, b, \beta}} + C_\rho \lambda_1 + \lambda_1^r \|g_\rho\|_{L^2_{\rho_X}}$. 结论 (2.2) 式成立.

参 考 文 献

- [1] Smale S, Zhou D X. Online learning with Markov sampling[J]. Analysis and Applications, 2009, 7: 87–113.
- [2] Steinwart I, Hush D, Scovel C. Learning from dependent observations[J]. Journal of Multivariate Analysis, 2009, 100: 175–194.
- [3] Hu T, Zhou D X. Online learning with samples drawn from non-identical distributions[J]. Journal of Machine Learning Research, 2009, 10: 2873–2898.
- [4] Koenker R, Bassett G. Regression quantiles[J]. Econometrica, 1978, 46: 33–50.
- [5] Koenker R. Quantile regression[M]. London: Cambridge University Press, 2005.

- [6] Steinwart I, Christmann A. How support vector machines can estimate quantiles and the median[J]. Advances in Neural Information Processing Systems, 2008, 20: 305–312.
- [7] Aronszajn N. Theory of reproducing kernels[J]. Transactions of the American Mathematical Society, 1950, 68: 337–404.
- [8] Ying Y, Zhou D X. Online regularized classification algorithms[J]. IEEE Transactions on Information Theory, 2006, 52: 4775–4788.
- [9] Hu T, Xiang D H, Zhou D X. Online learning for quantile regression and support vector regression[J]. Journal of Statistical Planning and Inference, 2012, 142: 3107–3122.
- [10] Smale S, Zhou D X. Learning theory estimates via integral operators and their applications[J]. Constructive Approximation, 2007, 26: 153–172.

QUANTILE REGRESSION WITH SAMPLES DRAWN FROM NON-IDENTICAL DISTRIBUTIONS

YANG Peng-wei

(*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*)

Abstract: This paper considers the online quantile regression with Gaussain kernel and non-identical sampling process. Under the sparsity condition and non-identical marginal distributions, we derive the convergence rate of the the online quantile regression algorithm. Mathematical analysis depends on the error decomposition and the iteration method. Specially, we get the explicit learning rate of online median regression. Finally, we illustrate that our main result can extend to general quantile regression problem.

Keywords: online learning; quantile regression; ϵ -pinball loss; reproducing kernel Hilbert space, non-identical distribution

2010 MR Subject Classification: 62J99