arXiv:2302.00975v1 [math.ST] 2 Feb 2023

Wasserstein 距离下分布回归的 Stone 定理

Clément Dombry[1], Thibault Modeste[2], 和 Romain Pic[1]

[1]弗朗什-孔泰大学，法国国家科学研究中心 UMR 6623，贝桑松数学实验室，F-25000 贝桑松，法国[2]里昂第一大学，法国国家科学研究中心 UMR 5208，卡米耶·若尔当研究所，F-69622 维勒班，法国

摘要

我们将著名的 Stone 定理推广到分布回归框架。具体而言，我们证明，采用满足 Stone 定理条件的局部概率权重构造的加权经验分布，可以对条件分布给出普遍一致估计，其中误差以 $p \geq 1$ 阶 Wasserstein 距离衡量。此外，对于 $p = 1$，我们确定了特定分布类别上的极小极大收敛速率。最后，我们给出这些结果的一些应用，包括条件尾部期望或概率加权矩的估计。

关键词： 分布回归、Wasserstein距离、非参数回归、极小极大收敛速率。

2020年数学主题分类：62G05。

目录

1 引言 2

2 背景 **3**
2.1 Stone 定理.32.2 Wasserstein 空间.6

3 主要结果 **7**
3.1 分布回归的 Stone 定理.73.2 收敛速率.73.3 应用.10

4 证明 **13**
4.1 定理2 的证明.134.2 命题1、推论1-2 和定理3 的证明 194.3 命题2 的证明.24

1 引言

预测是统计学的一项重要任务，且往往对决策制定至关重要。在最简单的情形下，当感兴趣的量是单变量且为定量变量时，点预测通常采用回归的形式，旨在估计响应变量 Y 在给定由协变量向量 X 所编码的可用信息的条件下的条件均值（或条件分位数）。点预测只是一个粗略的汇总统计量，至少应附带不确定性评估（例如标准差或置信区间）。或者，概率预测和分布回归（Gneiting和Katzfuss，2014）建议估计 Y 在给定 X 条件下的完整条件分布，称为预测分布。

近几十年来，天气预报一直是推动概率预测发展的重要动力。集合预报基于一定数量的确定性模型，这些模型的参数略有不同，以便将观测误差和大气不完整的物理表征纳入考量。这会产生一组不同的预测结果，同时也对预测的不确定性进行总体评估。集合预报存在偏差和离散不足（Hamill和Colucci，1997年），需要进行统计后处理以加以改进。人们提出了多种后处理方法，例如集合模型输出统计（Gneiting等，2005）、分位数回归森林（Taillardat等，2019）或神经网络（Schulz和Lerch，2021年）等。分布回归如今已广泛应用于气象学以外的领域，近期的方法学研究包括由Li等（2021年）提出的深度分布回归、分布

随机森林由Ćevid 等人（2022年）提出，或由Henzi 等人（2021年）提出的保序分布回归。

本文旨在将著名的 Stone 定理（Stone，1977年）推广到分布回归框架中。Stone 定理指出，局部权重算法能够一致地估计回归函数。Stone 定理的优势在于它完全非参数且不依赖模型，所需假设非常宽松，涵盖了许多重要情形，例如核算法和近邻方法。更多详情请参见Györfi 等人（2002年）。我们证明，Stone 定理可以自然而优雅地推广至分布回归，其中误差由阶数为 $p \geq 1$的 Wasserstein 距离度量。我们的结果不仅涵盖一维输出 $Y \in \mathbb{R}$ 的情形（此时 Wasserstein 距离具有简单的显式形式），也涵盖多维输出 $Y \in \mathbb{R}^d$的情形。采用 Wasserstein 距离的动机来自近期研究，这些研究表明它是统计学中一种实用且强大的工具，相关综述可参见 Panaretos 和 Zemel（2020年）。除这一主要结果外，我们还刻画了在适当的分布类别上，$d = 1$ 和 $p = 1$情形下的最优极小极大收敛速率。我们还讨论了这些结果对于估计若干可能感兴趣的统计量所具有的意义，例如预期短缺或概率加权矩。

本文结构如下。在第 2 节中，我们介绍 Stone 定理和 Wasserstein 空间所需的背景知识。第 3 节汇集了我们的主要结果，包括将 Stone 定理推广到分布回归（定理 2）、对最优极小极大收敛速率的刻画（定理 3）以及一些应用（命题 2及其后续示例）。所有技术性证明均推迟至第 4 节。

2 背景

2.1 Stone 定理

在回归框架中，我们观察到一个样本 (X_i, Y_i)，$1 \leq i \leq n$，它由 $(X, Y) \in \mathbb{R}^k \times \mathbb{R}^d$ 的独立副本构成，且其分布为 P。基于该样本并假设 Y 可积，目标是估计回归函数

$$r(x) = \mathbb{E}[Y|X = x], \quad x \in \mathbb{R}^k.$$

局部平均估计量的形式为

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \tag{1}$$

其中 $W_{n1}(x) \ldots, W_{nn}(x)$ 是 x 处的局部权重 。假设局部权重是 x 和 $X_1, \ldots, X_n$ 的可测函数，但不依赖于

$Y_1,\ldots,Y_n$，即

$$W_{ni}(x) = W_{ni}(x; X_1, \ldots, X_n), \quad 1 \le i \le n. \tag{2}$$

为方便记号，隐含了对 $X_1,\ldots,X_n$ 的依赖。本文仅关注满足以下条件的概率权重 s

$$W_{ni}(x) \ge 0,\ 1 \le i \le n, \quad \text{and} \quad \sum_{i=1}^{n} W_{ni}(x) = 1. \tag{3}$$

Stone 定理指出，回归估计量在 L^p-范数下具有普遍一致性。

定理 1 (Stone(1977年))。 假设概率权重 (3) 满足以下三个条件：

i) 存在 $C > 0$，使得 $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) g(X_i)\right] \le C\mathbb{E}[g(X)]$ 对所有 $n \ge 1$ 以及可测的 g： $\mathbb{R}^k \to [0, +\infty)$ 满足条件 $\mathbb{E}[g(X)] < \infty$;

ii) for all $\varepsilon > 0$, $\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > \varepsilon\}} \to 0$ *in probability as* $n \to +\infty$;

iii) 最大值$_{1 \le i \le n} W_{ni}(X) \to 0$ 依概率收敛，当 $n \to +\infty$。

那么，对于所有满足 $\mathbb{E}[\|Y\|^p] < \infty$ 的 $p \ge 1$ 和 $(X,\ Y) \sim P$，

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \longrightarrow 0 \quad as\ n \to +\infty. \tag{4}$$

反过来，如果等式 (4) 成立，那么概率权重必须满足条件 $i) - iii)$。

注 1。 Stone 定理通常是在维度 $d = 1$下表述的。由于随机向量 $\hat{r}_n(X) \to r(X)$ 在 L^p 中的收敛等价于所有分量在 L^p 中的收敛，因此推广到维度$d \ge 2$是显而易见的。此外，还可以考虑比概率权重更一般的权重：条件 (3) 可以去掉，并以以下较弱的假设取代：

$$|W_{ni}(X)| \le M \quad \text{a.s. for some } M > 0.$$

and

$$\sum_{i=1}^{n} W_{ni}(X) \to 1 \text{ in probability.}$$

本文不考虑此类一般权重，因此我们将仅讨论概率权重。读者可参阅 Biau andDevroye(2015年)，其中给出了 Stone 定理的完整证明及相关讨论。

例 1. 以下两个核权重和最近邻权重的例子，是文献中最重要的例子。关于更多细节，请参阅Györfi 等人 (2002年) 第 5 章和第 6 章。

- 核权重定义为

$$W_{ni}(x) = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h_n}\right)}, \quad 1 \le i \le n \tag{5}$$

如果分母非零，则 $1/n$ 否则。这里，带宽 $h_n > 0$ 仅取决于样本大小， n而函数K： $\mathbb{R}^k \to [0, +\infty)$称为核。在这种情况下，估计量（1）对应于回归函数（Nadaraya，1964；Watson，1964）的 Nadaraya-Watson 估计量。我们称 K 为盒式核，如果存在常数 $R_2 \ge R_1 > 0$ 和 $M_2 \ge M_1 > 0$ 使得

$$M_1 \mathbb{1}_{\{\|x\| \le R_1\}} \le K(x) \le M_2 \mathbb{1}_{\{\|x\| \le R_2\}}, \quad x \in \mathbb{R}^k.$$

Györfi 等人的 (2002年) 定理 5.1 表明，对于盒式核，核权重（5）满足定理 $i) - iii)$的条件 1 当且仅当 $h_n \to 0$ 以及 $nh_n^k \to +\infty$ 如 $n \to +\infty$。

- 最近邻（NN）权重定义为

$$W_{ni}(x) = \begin{cases} \frac{1}{\kappa_n} & \text{if } X_i \text{ belongs to the } \kappa_n\text{-NN of } x \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

其中邻居数 $\kappa_n \in \{1, \dots, n\}$ 仅取决于样本大小。回顾一下，样本内 κ_n-NN x 的 $(X_i)_{1 \le i \le n}$是通过按距离 $\|X_i - x\|$ 递增排序，并保留距离最小的 κ_n 个点获得的——正如 Györfi 等人 (2002年) 第 6 章所述，可以使用多种规则来打破平局，例如字典序或随机破平局。同一参考文献中的定理 6.1 表明，最近邻权重（6）满足定理 $i) - iii)$的条件 1 当且仅当 $\kappa_n \to +\infty$ 以及 $\kappa_n/n \to 0$ 如 $n \to +\infty$。

例 2. 有趣的是，著名的 Breiman 随Breiman，2001机森林的某些变体会产生满足 Stone 定理假设的概率权重。在 Breiman 随机森林中，分裂同时涉及协变量和响应变量，因此相应的权重 $W_{ni}(x) = W_{ni}(x; (X_l, Y_l)_{1 \le l \le n})$ 并非形式（2）。Scornet(2016）考察了无限随机森林的两个简化版本，其中相应的权重 $W_{ni}(x)$ 不依赖于响应值，并满足所谓的 X-性质，即它们具有形式（2）。对于完全

非自适应森林中，树通过一种不使用训练样本且完全随机的二元分裂规则生长；作者证明，在叶节点数量以小于 n的速率趋于无穷且叶节点体积依概率趋于零的条件下，与无限森林相关的概率权重满足 Stone 定理的假设（见定理 4.1 及其证明）。对于 q-分位数森林，二元分裂规则仅涉及协变量；作者证明，只要子抽样数量 a_n 满足 $a_n \to +\infty$ 且$a_n/n \to 0$，与无限森林相关的权重就满足 Stone 定理的假设（见定理 5.1 及其证明）。

2.2 Wasserstein 空间

我们回顾 $\mathbb{R}^d$ 上 Wasserstein 空间的定义及一些基本事实。关于最优传输和 Wasserstein 空间的更多细节及进一步结果，可参见 Villani (2009) 的专著第 6 章。

对于 $p \geq 1$，Wasserstein 空间 $\mathcal{W}_p(\mathbb{R}^d)$ 定义为 $\mathbb{R}^d$ 上具有 p 阶有限矩的博雷尔概率测度集合，即满足

$$M_p(\mu) = \Big(\int_{\mathbb{R}^d} \|y\|^p \, \mu(\mathrm{d}y) \Big)^{1/p} < \infty. \tag{7}$$

其赋予如下定义的距离：对于 Q_1， $Q_2 \in \mathcal{W}_p(\mathbb{R}^d)$，

$$\mathcal{W}_p(Q_1, Q_2) = \inf_{\pi \in \Pi(Q_1, Q_2)} \left(\int \|y_1 - y_2\|^p \, \pi(\mathrm{d}y_1 \mathrm{d}y_2) \right)^{1/p}, \tag{8}$$

其中 $\Pi(Q_1, Q_2)$ 表示 $\mathbb{R}^d \times \mathbb{R}^d$ 上边缘分布为 Q_1 和 Q_2 的测度集合。一对 (Z_1, Z_2) 随机变量，其分布分别为 Q_1 和 Q_2，称为耦合。因此，Wasserstein 距离是所有可能耦合中距离 $\|Z_1 - Z_2\|_{\mathrm{L}^p} = \mathbb{E}[\|Z_1 - Z_2\|^p]^{1/p}$ 的最小值。由于 $\mathbb{R}^d$ 是完备可分度量空间，最优耦合的存在性得以保证，因此下确界确实是最小值。

Wasserstein 距离通常很难计算，但情形$d = 1$ 是个例外。概率逆变换给出了一种简单的最优耦合：对于 $i = 1, 2$，令 $Q_i \in \mathcal{W}_p(\mathbb{R})$， F_i 表示其累积分布函数， F_i^{-1}表示其广义逆函数（分位数函数）。然后，从均匀随机变量 $U \sim \mathrm{Unif}(0, 1)$ 出发，可得一个最优耦合 $(Z_1, Z_2) = (F_1^{-1}(U)\ F_2^{-1}(U))$。因此，Wasserstein 距离可明确写为

$$\mathcal{W}_p(Q_1, Q_2) = \left(\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p \mathrm{d}u \right)^{1/p}. \tag{9}$$

当 $p = 1$ 时，简单的变量替换可得

$$\mathcal{W}_1(Q_1, Q_2) = \int_{-\infty}^{+\infty} |F_1(u) - F_2(u)| \mathrm{d}u. \tag{10}$$

3 主要结果

3.1 分布回归的 Stone 定理

现在我们给出本文的主要结果，它将 Stone 定理自然地推广到分布回归框架。给定定义在 $\mathbb{R}^k \times \mathbb{R}^d$ 上的分布 $(X,Y) \sim P$，我们将 F 记为 Y 的边缘分布，将 F_x 记为给定 $X = x$ 时的条件分布。该条件分布可以通过独立同分布于 (X,Y) 的样本 $(X_i,Y_i)_{1\le i\le n}$，利用加权经验分布进行估计

$$\hat{F}_{n,x} = \sum_{i=1}^{n} W_{ni}(x)\delta_{Y_i} \tag{11}$$

其中 δ_y 表示点 $y \in \mathbb{R}^d$处的狄拉克测度。对于满足（ˆ3）的概率权重，$F_{n,x}$ 是一个概率测度，并可视为完备可分空间 $\mathcal{W}_p(\mathbb{R}^d)$ 中的随机元素。我们回顾一下，权重 $W_{ni}(x) = W_{ni}(x; X_1,\dots,X_n)$ 隐式依赖于 $X_1,\dots,X_n$，但不依赖于 $Y_1,\dots,Y_n$。

定理 2。 假设概率权重满足定理 $i) - iii)$ 中的条件 *1*。那么，对于所有 $p \ge 1$ 以及 (X,Y) 满足 $\mathbb{E}[\|Y\|^p] < \infty$的情形，

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] \longrightarrow 0 \quad as\ n \to +\infty. \tag{12}$$

反之，若等式 (12) 成立，则概率权重必须满足条件 $i) - iii)$。

值得注意的是

$$\mathbb{E}\left[\|\hat{r}_n(X) - r(X)\|^p\right] \le \mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right]$$

因此，定理 2 可直接推出定理 1 。定理 2 的证明推迟到第 4 节。首先考虑以下情形： $d = 1$ Wasserstein 距离由公式（9）明确给出。随后，借助最大切片 Wasserstein 距离的概念（Bayraktar 和 Guo，2021年），将结果推广到更高维度 $d \ge 2$ ；该概念使得可以将 $\mathbb{R}^d$ 上测度的收敛性归结为其一维投影的收敛性（定理 4 给出了下文中的精确表述）。

3.2 收敛速度

接下来，我们考虑极小极大意义下的收敛速率。请注意，类似的问题和结果已在 Pic等人 (2022年中得到确立，其中考虑了二阶Cramér距离，即

$$\|\hat{F}_{n,X} - F_X\|_{L_2}^2 = \int_{\mathbb{R}} |\hat{F}_{n,X}(y) - F_X(y)|^2 \,\mathrm{d}y.$$

ˆ这里我们关注 Wasserstein 距离 $\mathcal{W}_p(F_{n,\ X},\ F_X)$，并且仅考虑 $d=1$ 和 $p=1$ 这两种情况，因为它们允许得到显式表达式（10）。其他情况似乎更难分析，超出了本文的研究范围。我们的第一个结果考察了当 $X=x$ 固定时 Wasserstein 距离上的误差。

命题 1. 假设 $d=1$ *and* $(X,Y)\sim P$满足 $\mathbb{E}[|Y|]<\infty$。那么，

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,x},F_x)\big]\le\mathbb{E}\Big[\sum_{i=1}^n W_{ni}(x)\mathcal{W}_1(F_{X_i},F_x)\Big]+M(x)\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(x)\Big]^{1/2},$$

其中 $M(x)=\int_{\mathbb{R}}\sqrt{F_x(z)(1-F_x(z))}\mathrm{d}z$。

第一项对应于近似误差，这是因为我们使用有偏样本来估计 F_x。模型越规则，近似误差就越小。第二项是估计误差，这是因为我们使用经验均值来估计 F_x。如果分布误差的离散度较低（用 $M(x)$ 衡量），或者 $\sum_{i=1}^n W_{ni}^2(x)$ 较小，那么该估计误差就会更小。请注意，对于最近邻权重，$1/\sum_{i=1}^n W_{ni}^2(x)$ 恰好等于 κ，因此这一数量通常称为有效样本量，而估计误差与期望有效样本量倒数的平方根成正比。

鉴于命题 1，我们引入以下函数类。

定义 1. 令 $\mathcal{D}(H,\ L,\ M)$ 为满足以下条件的分布类$(X,\ Y)\sim P$于$\mathbb{R}^k\times\mathbb{R}$：

a) $X\in[0,1]^k$ *a.s. and* $\mathbb{E}|Y|<\infty$,

b) for all $x,x'\in[0,1]^k$, $\mathcal{W}_1(F_x,F_{x'})\le L\|x-x'\|^H$,

c) for all $x\in[0,1]^k$, $\int_{\mathbb{R}}\sqrt{F_x(z)(1-F_x(z))}\,\mathrm{d}z\le M$.

该类别的定义与命题 1 意味着，类别 $\mathcal{D}(H,\ L,\ M)$ 上的期望误差一致有界，其界为

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{W}_1(\hat{F}_{n,X},F_X)\Big]\\&\le L\mathbb{E}\Big[\sum_{i=1}^n W_{ni}(X)\|X_i-X\|^H\Big]+M\mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(X)\Big]^{1/2}.\end{aligned}\tag{13}$$

因此，命题1 使我们能够针对核方法和最近邻方法，从例1中推导出关于 $\mathcal{D}(H,L,M)$ 的一致显式界。为简洁起见，我们仅考虑均匀核。

ˆ**推论1**。设 $F_{n,X}$ 由采用均匀核$K(x) = \mathbb{1}_{\{\|x\|\leq 1\}}$的核方法给出，且权重由方程(5)给出。若 $P \in \mathcal{D}(H, L, M)$，则

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] \leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}$$

与 $c_k = k^{k/2}$。

ˆ**推论2**。令 $F_{n,X}$ 由最近邻方法给出，其权重由公式 (6) 给出，并假设 $P \in \mathcal{D}(H, L, M)$。于是，

$$\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] \leq \begin{cases} L8^{H/2}(\kappa_n/n)^{H/2} + M\kappa_n^{-1/2} & \text{if } k = 1, \\ L\tilde{c}_k^{H/2}(\kappa_n/n)^{H/k} + M\kappa_n^{-1/2} & \text{if } k \geq 2, \end{cases}$$

其中 $\tilde{c}_k$ 仅取决于维数 k，定义见 *Biau and Devroye(2015年, 定理2.4)*。

可以看出，只要 $h_n \to 0$ 且 $nh_n^k \to +\infty$ （对于核方法），以及 $\kappa_n/n \to 0$ 且 $\kappa_n \to +\infty$ （对于最近邻方法），一致性就成立——也就是说，当 $n \to +\infty$ 时，期望误差趋于 0。

下一个定理给出了类别 $\mathcal{D}(H, L, M)$ 上的最优极小极大收敛速率。我们称两个正数序列 (a_n) 和 (b_n) 具有相同的收敛速率，记作 $a_n \asymp b_n$，如果比值 a_n/b_n 和 b_n/a_n 在 $n \to +\infty$ 时保持有界。

定理 3. 类别 $\mathcal{D}(H, L, M)$上的最优极小极大收敛速率为

$$\inf_{\hat{F}_n} \sup_{P \in \mathcal{D}(H,L,M)} \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)] \asymp n^{-H/(2H+k)}.$$

定理 3是 Pic等人 (2022年, 定理 1) 的对应结果，其中考察了二阶 Cramér距离的极小极大收敛速率。证明策略类似：i) 我们考虑一个合适的二元分布类别，证明下界；在该类别中，Wasserstein 距离上的误差对应于点回归中的绝对误差，而后者的极小极大下界收敛速率已知；ii) 我们验证核方法和/或最近邻算法的上界具有与下界相同的收敛速率，从而证明已确定最优极小极大收敛速率。具体而言，我们的证明表明，式（5）中定义的核方法在任意维度下都能达到极小极大收敛速率 $k \geq 1$，其带宽取为 $h_n \asymp n^{-1/(2H+k)}$；式（6）中定义的最近邻方法在任意维度下都能达到极小极大收敛速率$k \geq 2$，其邻居数取为 $\kappa_n \asymp n^{H/(H+k/2)}$。

备注 2. 我们的极小极大收敛速率估计仅适用于$d = p = 1$，下面简要讨论其他情形下可能出现的情况。

当 $p = 1$ 且 $d \geq 2$时，人们或许可以利用最大切片 Wasserstein 距离与 Wasserstein 距离之间的强等价关系（Bayraktar 和 Guo，2021年，定理 2.3.ii）。这需要估计球面上的上确界的期望，这一研究方向留待后续工作。

当 $p > 1$时，即使在维度为 $d = 1$的情况下，若没有非常强的假设，似乎也难以获得 Wasserstein 距离的 p 阶界。Bobkov 和 Ledoux（2019）研究了独立同分布样本 $Y_1,\dots,Y_n$ 的经验分布 $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i}$ 的收敛速率，其分布为 F，定义在 $\mathbb{R}$ 上。首个一致性结果（定理 2.14）指出，只要 F 具有 $p \geq 1$阶有限阶矩，就有 $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_n, F)] \to 0$。关于收敛速率，他们证明（推论 3.9），对于 $p = 1$，$\hat{}$ 的标准收敛速率成立，即 $\mathbb{E}[\mathcal{W}_1(F_n, F)] = O(1/\sqrt{n})$，当且仅当

$$J_1(F) = \int_{\mathbb{R}} \sqrt{F(z)(1-F(z))}\mathrm{d}z < \infty.$$

另一方面，更高阶 $p > 1$ 的收敛速率需要满足条件

$$J_p(F) = \int_{\mathbb{R}} \frac{[F(z)(1-F(z))]^{p/2}}{f(z)^{p-1}}\mathrm{d}z < \infty,$$

其中，f 是 F 的绝对连续部分的密度。他们证明（推论 5.5），当且仅当 $J_p(F) < \infty$ 时，标准速率成立，即 $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_n, F)] = O(n^{-p/2})$。然而，这个条件非常严格：对于高斯分布或支撑不连通的分布，该条件并不成立。

3.3 应用

我们简要介绍定理2 的一些应用和示例。在统计学中，我们通常会遇到以下一般情形：我们关注一个取实数值的汇总统计量 S，例如分位数或尾部期望，并希望评估 X 对 Y 的影响，具体而言，通过 S，也就是说，我们希望评估 $S_{Y|X=x}$。假设 S 对于定义在 $\mathbb{R}^d$ 上且具有 $p \geq 1$阶有限矩的分布是良定义的，那么它可以看作一个映射 $S:\ \mathcal{W}_p(\mathbb{R}^d) \to \mathbb{R}$，进而 $S_{Y|X=x} = S(F_x)$，其中 F_x 是给定$X = x$时 Y 的条件分布。对 $S_{Y|X=x}$ 的一种自然代入估计量是

$$\hat{S}_{n,x} = S(\hat{F}_{n,x}) \quad \text{with } \hat{F}_{n,x} \text{ defined by (11).}$$

在这一一般情形下，我们对 Stone 定理的推广直接推出以下命题。回想一下，$M_p(\mu)$在方程（7）中定义。

命题 2. 假设 $\mathbb{E}[\|Y\|^p] < \infty$ 和 $\mathbb{P}(F_X \in \mathcal{C}) = 1$，其中 $\mathcal{C} \subset \mathcal{W}_p(\mathbb{R}^d)$表示统计量 $S: \mathcal{W}_p(\mathbb{R}^d) \to \mathbb{R}$的连续性集合。那么弱一致性成立，即

$$\hat{S}_{n,X} \longrightarrow S_{Y|X} \quad in\ probability\ as\ n \to +\infty.$$

此外，若统计量 S 满足如下形式的界

$$|S(\mu)| \leq aM_p^q(\mu) + b, \quad with\ a, b \geq 0\ and\ 0 < q \leq p, \tag{14}$$

则一致性在$\mathrm{L}^{p/q}$中成立，即

$$\mathbb{E}\big[|\hat{S}_{n,X} - S_{Y|X}|^{p/q}\big] \longrightarrow 0 \quad as\ n \to +\infty$$

例 3. </style》（分位数）。对于定义在 $\mathbb{R}$ 上的分布 G，我们定义相应的分位数函数

$$G^{-1}(\alpha) = \inf\{z \in \mathbb{R} : G(z) \geq \alpha\}, \quad 0 < \alpha < 1.$$

众所周知，弱收敛 $G_n \xrightarrow{d} G$ 蕴含在 G^{-1} 的每个连续点 α 处，分位数收敛 $G_n^{-1}(\alpha) \to G^{-1}(\alpha)$。等价地，考虑赋予弱收敛拓扑的 $\mathcal{P}(\mathbb{R})$，只要 G^{-1} 在 α 处连续，α 分位数统计量 $S_\alpha(G) = G^{-1}(\alpha)$ 就在 G 处连续。

鉴于此，我们令 $\mathcal{C} = \{G \in \mathcal{P}(\mathbb{R})$ 在 $(0, 1)$ 上连续，并假设条件分布满足 $\mathbb{P}(F_X \in \mathcal{C}) = 1$。那么条件分位数弱收敛，即

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in probability.}$$

注意，此处不需要可积性条件，因为我们可以对变换数据 2 应用命题 $(X_i, \tilde{Y}_i)_{1 \leq i \leq n}$，其中 $\tilde{Y}_i = \tan^{-1}(Y_i)$有界，因此 Wasserstein 距离收敛等价于弱收敛。此外，若 Y 满足 p-可积，则有界

$$\begin{aligned} |S_\alpha(G)|^p &\leq \frac{1}{\alpha}\int_0^\alpha |G^{-1}(u)|^p \mathrm{d}u + \frac{1}{1-\alpha}\int_\alpha^1 |G^{-1}(u)|^p \mathrm{d}u \\ &\leq \Big(\frac{1}{\alpha} + \frac{1}{1-\alpha}\Big) M_p^p(G) \end{aligned}$$

意味着更强的收敛

$$\hat{F}_{n,X}^{-1}(\alpha) \to F_X^{-1}(\alpha) \quad \text{in } \mathrm{L}^p.$$

例 4.（尾部期望）阈值 $\alpha \in (0, 1)$ 以上的尾部期望是为 $G \in \mathcal{W}_1(\mathbb{R})$ 定义的风险度量，定义如下：

$$S_\alpha(G) = \frac{1}{1-\alpha}\int_\alpha^1 G^{-1}(u)\,\mathrm{d}u.$$

这个名称源自其等价定义

$$S_\alpha(G) = \mathbb{E}[Y \mid Y > G^{-1}(\alpha)], \quad Y \sim G,$$

当 G^{-1} 在 α处连续时，该式成立。可以看出

$$\begin{aligned} |S_\alpha(G_1) - S_\alpha(G_2)| &\leq \frac{1}{1-\alpha}\int_\alpha^1 |G_1^{-1}(u) - G_1^{-1}(u)|\,\mathrm{d}u \\ &\leq \frac{1}{1-\alpha}\int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|\,\mathrm{d}u \\ &= \frac{1}{1-\alpha}\mathcal{W}_1(G_1, G_2). \end{aligned}$$

因此， S_α 关于 Wasserstein 距离$\mathcal{W}_1$是 Lipschitz 连续的。因此，由于以下关系，条件尾部期望 $S_\alpha(F_x)$ 可以通过代入估计量 $S_\alpha(F_{n,\ x})$ 一致地估计

$$\mathbb{E}[|S_\alpha(\hat{F}_{n,X}) - S_\alpha(F_X)|] \leq \frac{1}{1-\alpha}\mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)] \longrightarrow 0.$$

例 5。（概率加权矩）对于阶数为 p、$q > 0$ 的概率加权矩，也有类似结果，其定义为

$$S_{p,q}(G) = \int_0^1 G^{-1}(u)u^p(1-u)^q\,\mathrm{d}u, \quad G \in \mathcal{W}_1(\mathbb{R}).$$

（Greenwood 等人 （1979）。其名称源于等价定义

$$S(G) = \mathbb{E}[YG(Y)^p(1-G(Y))^q], \quad Y \sim G,$$

当 G^{-1} 在 (0, 1) 上连续时，该式成立。同样可以验证，由于以下关系，统计量 $S_{p,\ q}$ 关于 Wasserstein 距离$\mathcal{W}_1$ 是 Lipschitz 连续的

$$\begin{aligned} |S_{p,q}(G_1) - S_{p,q}(G_2)| &\leq \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|u^p(1-u)^q\,\mathrm{d}u \\ &\leq \max_{0\leq u\leq 1} u^p(1-u)^q \times \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|\,\mathrm{d}u \\ &= \Big(\frac{p}{p+q}\Big)^p\Big(\frac{q}{p+q}\Big)^q\mathcal{W}_1(G_1, G_2). \end{aligned}$$

例 6.（协方差）最后，我们给出一个简单的例子，讨论维度为 $d = 2$ 的情形，其中感兴趣的统计量是给定 $X = x$ 时 $Y =$ （Y_1, Y_2）两个分量之间的协方差。在此，我们考虑

$$S(G) = \int_{\mathbb{R}^2} y_1y_2\,\mathrm{d}G - \int_{\mathbb{R}^2} y_1\,\mathrm{d}G\int_{\mathbb{R}^2} y_2\,\mathrm{d}G, \quad G \in \mathcal{W}_2(\mathbb{R}^2).$$

考虑分布分别为 G 和 H 的平方可积随机向量 $Y = (Y_1, Y_2)$ 和 $Z = (Z_1, Z_2)$，我们计算

$$\begin{aligned}
&|S(G) - S(H)| \\
&= \big|\mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(Z_1, Z_2)\big| \\
&= \big|\mathrm{Cov}(Y_1, Y_2 - Z_2) - \mathrm{Cov}(Z_1 - Y_1, Z_2)\big| \\
&\le \mathrm{Var}(Y_1)^{1/2}\mathrm{Var}(Y_2 - Z_2)^{1/2} + \mathrm{Var}(Z_2)^{1/2}\mathrm{Var}(Z_1 - Y_1)^{1/2}
\end{aligned}$$

其中最后一行由柯西-施瓦茨不等式得出。我们得到上界

$$\mathrm{Var}(Y_1)^{1/2} \le M_2(G), \quad \mathrm{Var}(Z_2)^{1/2} \le M_2(H)$$

并且，选取 G 与 H之间的一个最优耦合 (Y, Z),

$$\mathrm{Var}(Z_1 - Y_1)^{1/2} \le \|Y - Z\|_{L^2} = \mathcal{W}_2(G, H), \quad \mathrm{Var}(Y_2 - Z_2)^{1/2} \le \mathcal{W}_2(G, H).$$

综上，我们得到:

$$|S(G) - S(H)| \le \big(M_2(G) + M_2(H)\big)\mathcal{W}_2(G, H).$$

这证明 S 是局部利普希茨的，因此相对于距离 $\mathcal{W}_2$连续。取 $H = \delta_0$，我们得到

$$|S(G)| \le M_2(G)^2$$

且界（14）在 $q = 2$ 时成立。因此，命题 2 表明，代入估计量

$$S(\hat{F}_{n,x}) = \sum_{i=1}^{n} W_{ni}(x) Y_{1i} Y_{2i} - \sum_{i=1}^{n} W_{ni}(x) Y_{1i} \sum_{i=1}^{n} W_{ni}(x) Y_{2i}$$

在绝对均值意义下是一致估计条件协方差的

$$S(F_x) = \mathbb{E}(Y_1 Y_2 \mid X = x) - \mathbb{E}(Y_1 \mid X = x)\mathbb{E}(Y_2 \mid X = x),$$

ˆ即 $\mathbb{E}[|S(F_{n,\ x}) - S(F_X)|] \longrightarrow 0$作为 $n \to +\infty$。

4 证明

4.1 定理的证明 2

定理的证明 *2* - 情形 $d = 1$。我们首先考虑以下情形： Y一致有界，且其取值属于 $[-M, M]$，其中 $M > 0$。此时有

$$F_x(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \ge M \end{cases} \quad \text{and} \quad \hat{F}_{n,x}(z) = \begin{cases} 0 & \text{if } z < -M \\ 1 & \text{if } z \ge M \end{cases}.$$

并且广义逆函数（分位数函数）的绝对值均不超过 M。因此，

$$\begin{aligned}\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] &= \mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)|^p \mathrm{d}z\right] \\ &\leq (2M)^{p-1}\mathbb{E}\left[\int_0^1 |\hat{F}_{n,X}^{-1}(u) - F_X^{-1}(u)| \mathrm{d}u\right] \\ &= (2M)^{p-1}\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z. \qquad (15)\end{aligned}$$

在这些推导中，我们结合使用了方程（9）和（10）以及 Fubini定理。

考虑回归模型 $(X, \mathbb{1}_{\{Y\leq z\}}) \in \mathbb{R}^d \times \mathbb{R}$ 其中 $z \in [-M, M]$固定。相应的回归函数为

$$x \mapsto \mathbb{E}[\mathbb{1}_{\{Y\leq z\}}|X = x] = F_x(z)$$

与样本 $(X_i, \mathbb{1}_{\{Y_i\leq z\}})$,$1 \leq i \leq n$ 对应的局部权重估计量为

$$x \mapsto \sum_{i=1}^{n} W_{ni}(x)\mathbb{1}_{\{Y_i\leq z\}} = \hat{F}_{n,x}(z).$$

应用 Stone 定理并取 $p = 1$ ，可得

$$\mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \longrightarrow 0, \quad \text{as } n \to +\infty,$$

因此，根据控制收敛定理，我们推出

$$\int_{-M}^{M} \mathbb{E}\left[|\hat{F}_{n,X}(z) - F_X(z)|\right] \mathrm{d}z \longrightarrow 0.$$

上界（15）最终意味着

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, F_X)\right] \longrightarrow 0.$$

接下来考虑一般情形，即 Y 不一定有界。对于 $M > 0$，我们将 Y的截断 Y^M 定义为

$$Y^M = \begin{cases} -M & \text{if } Y < -M \\ Y & \text{if } -M \leq Y < M \\ M & \text{if } Y \geq M \end{cases}.$$

我们类似地定义 Y_1^M,... , Y_n^M分别为 $Y_1, \ldots, Y_n$ 的截断。与 Y^M 相关联的条件分布为

$$F_x^M(z) = \mathbb{P}(Y^M \leq z|X = x) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq Y < M \\ 1 & \text{if } z \geq M \end{cases}.$$

基于截断样本构建的局部权重估计为

$$\hat{F}_{n,x}^{M}(z)=\sum_{i=1}^{n}W_{ni}(x)\mathbb{1}_{\{Y_i^M\le z\}}.$$

由三角不等式,

$$\mathcal{W}_p(\hat{F}_{n,x},F_x)\le\mathcal{W}_p(\hat{F}_{n,x},\hat{F}_{n,x}^{M})+\mathcal{W}_p(\hat{F}_{n,x}^{M},F_x^{M})+\mathcal{W}_p(F_x^{M},F_x),$$

由此可得

$$\begin{aligned}&\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,x},F_x)]\\&\le 3^{p-1}\left(\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X},\hat{F}_{n,X}^{M})]+\mathbb{E}[\mathcal{W}_p(\hat{F}_{n,X}^{M},F_X^{M})]+\mathbb{E}[\mathcal{W}_p^p(F_X^{M},F_X)]\right).\end{aligned}$$

根据有界情形下的前述结果，对于任意固定的 M，第二项在 $n\to+\infty$ 时趋于 0。接下来我们关注第一项和第三项。

对于固定的 $X=x$，分布 $\hat{F}_{n,x}$ 与 $\hat{F}_{n,x}^{M}$ 之间存在一种自然的耦合，由 (Z_1,Z_2) 给出，使得

$$(Z_1,Z_2)=(Y_i,Y_i^M)\quad\text{with probability } W_{ni}(x).$$

显然，$Z_1\sim\hat{F}_{n,x}$ 和 $Z_2\sim\hat{F}_{n,x}^{M}$，且该耦合给出了上界

$$\mathcal{W}_p^p(\hat{F}_{n,x},\hat{F}_{n,x}^{M})\le\|Z_1-Z_2\|_{\mathrm{L}^p}^p=\sum_{i=1}^{n}W_{ni}(x)|Y_i-Y_i^M|^p.\tag{16}$$

引入由下式定义的函数 $g_M(x)$

$$g_M(x)=\mathbb{E}\left[|Y-Y^M|^p\mid X=x\right].$$

利用在给定 $X_1,\dots,X_n$ 的条件下，随机变量 $Y_1,\dots,Y_n$ 相互独立且服从分布 $F_{X_1},\dots,F_{X_n}$ 这一事实，我们推得

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,x},\hat{F}_{n,x}^{M})\right]\le\mathbb{E}\left[\sum_{i=1}^{n}W_{ni}(x)g_M(X_i)\right].$$

那么，Stone 定理中关于权重的条件 *i)* 蕴含

$$\mathbb{E}\left[\sum_{i=1}^{n}W_{ni}(X)g_M(X_i)\right]\le C\mathbb{E}[g_M(X)].$$

由于当 $M\to+\infty$ 时，$|Y-Y^M|^p$ 几乎必然收敛到 0，且被可积的 $2^p|Y|^p$ 所控制，根据勒贝格收敛定理可得

$$\mathbb{E}[g_M(X)]=\mathbb{E}\left[|Y-Y^M|^p\right]\longrightarrow 0\quad\text{as } M\to+\infty.$$

由此可推得第一项满足

$$\mathbb{E}\left[\mathcal{W}_p^p(\hat{F}_{n,X}, \hat{F}_{n,X}^M)\right] \leq C\mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

其中，收敛关于 n 是一致的。

现在考虑第三项。由于 Y^M 是通过截断从 Y 得到的，因此 Y 和 Y^M 的分布函数与分位数函数之间满足

$$F_x^M(z) = \begin{cases} 0 & \text{if } z < -M \\ F_x(z) & \text{if } -M \leq z < M \\ 1 & \text{if } z \geq M \end{cases}$$

and

$$(F_x^M)^{-1}(u) = \begin{cases} -M & \text{if } F_x^{-1}(u) < -M \\ (F_x)^{-1}(u) & \text{if } -M \leq F_x^{-1}(u) < M \\ M & \text{if } F_x^{-1}(u) \geq M \end{cases}.$$

因此

$$\begin{aligned} \mathcal{W}_p^p(F_x^M, F_x) &= \int_0^1 |(F_x^M)^{-1}(u) - F_x^{-1}(u)|^p \mathrm{d}u \\ &= \mathbb{E}\left[|Y^M - Y|^p \mid X = x\right] = g_M(x). \end{aligned}$$

我们推得

$$\mathbb{E}\left[\mathcal{W}_p^p(F_X^M, F_X)\right] = \mathbb{E}[g_M(X)] \longrightarrow 0, \quad \text{as } M \to +\infty$$

其中，收敛关于 n 是一致的。

最后，我们将这三项合并。通过先选择 M 足够大，使第一项和第三项小于 $\varepsilon/3$，再选择 n 足够大，使第二项小于 $\varepsilon/3$，可以使该和小于任意 $\varepsilon > 0$。这就证明了方程（12），并完成了证明。

□

为了将证明从 $d = 1$ 推广到 $d \geq 2$，我们需要引入切片 Wasserstein 距离的概念，例如参见 Bayraktar 和 Guo（2021）。令$\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d$ 为 $\mathbb{R}^d$ 中的单位球面，对于 $u \in \mathbb{R}^d$，令u_*为由 $u_*(x) = u \cdot x$定义的线性形式。测度 μ 在方向 u上的投影定义为推前 $\mu \circ u_*^{-1}$，它是 $\mathbb{R}$上的一个测度。不等式 $|u \cdot x| \leq \|x\|$意味着，对于所有 $\mu \in \mathcal{W}_p(\mathbb{R}^d)$ 和 $u \in \mathbb{S}^{d-1}$，都有 $\mu \circ u_*^{-1} \in \mathcal{W}_p(\mathbb{R})$。 μ与 $\nu \in \mathcal{W}_p(\mathbb{R}^d)$之间的切片 Wasserstein 距离和最大切片 Wasserstein 距离分别定义为

$$S\mathcal{W}_p(\mu, \nu) = \left(\int_{\mathbb{S}^{d-1}} \mathcal{W}_p^p(\mu \circ u_*^{-1}, \nu \circ u_*^{-1})\, \sigma(\mathrm{d}u)\right)^{1/p},$$

其中， σ 表示 $\mathbb{S}^{d-1}$ 上的均匀测度，并且

$$\overline{SW}_p(\mu,\nu) = \max_{u\in\mathbb{S}^{d-1}} \mathcal{W}_p(\mu\circ u_*^{-1}, \nu\circ u_*^{-1}).$$

通俗地说，切片 Wasserstein 距离和最大切片 Wasserstein 距离分别是 μ 与 ν 的投影之间所有一维 Wasserstein 距离的平均值和最大值。以下结果对我们的证明至关重要。

定理 4 (Bayraktar 和 Guo (2021年))。 对于所有 $p \geq 1$， $S\mathcal{W}_p$ 和 $\overline{SW}_p$都是 $\mathcal{W}_p$ $(\mathbb{R}^d)$ 上的距离，且它们与 $\mathcal{W}_p$等价，即对于所有序列μ， μ_1， μ_2， $\ldots \in \mathcal{W}_p(\mathbb{R}^d)$

$$S\mathcal{W}_p(\mu_n,\mu)\to 0 \iff \overline{SW}_p(\mu_n,\mu)\to 0 \iff \mathcal{W}_p(\mu_n,\mu)\to 0.$$

定理 *2* 的证明——情形 $d \geq 2$。 为清晰起见，我们将证明分为三个步骤：

1）我们证明该结果在最大切片 Wasserstein 距离下成立，即$\mathbb{E}[\overline{SW}_p^p(\hat{F}_{n,\ X}, F_X)] \to 0$;

2）我们推导出 $\mathcal{W}_p(\hat{F}_{n,\ X}, F_X) \to 0$以概率成立;

3) 我们证明序列 $\mathcal{W}_p^p(\hat{F}_{n,\ X}, F_X)$ 是一致可积的。

第 2) 点和第 3) 点共同推出 $\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,\ X}, F_X)] \to 0$，这正是所需的。

步骤 1)。对于所有 $u \in \mathbb{S}^{d-1}$， 投影 $\hat{F}_{n,\ X}\circ u_*^{-1}$ 是加权经验分布

$$\hat{F}_{n,X}\circ u_*^{-1} = \sum_{i=1}^n W_{ni}(X)\delta_{Y_i\cdot u}.$$

将定理 2 应用于 1 维样本 $(Y_i\cdot u)_{i\geq 1}$ 可得

$$\mathbb{E}[\mathcal{W}_p^p(\hat{F}_{n,X}\circ u_*^{-1}, F_X\circ u_*^{-1})] \longrightarrow 0. \tag{17}$$

注意， $\mathbb{E}[|Y|^p] < \infty$ 蕴含 $\mathbb{E}[|Y\cdot u|^p] < \infty$，且 $Y\cdot u$ 的条件分布是 Y 的推前，即 $\mathcal{L}(Y\cdot u \mid X)$ $=F_X\circ u_*^{-1}$。

接下来我们考虑最大切片 Wasserstein 距离。沿方向 $u \in \mathbb{S}^{d-1}$ 的正则性会很有用。我们回顾一下，投影之间的 Wasserstein 距离以 Lipschitz 方式依赖于方向。更确切地说，根据 Bayraktar 和 Guo (2021年，命题 2.2)，

$$|\mathcal{W}_p(\mu\circ u_*^{-1},\nu\circ u_*^{-1}) - \mathcal{W}_p(\mu\circ v_*^{-1},\nu\circ v_*^{-1})| \leq (M_p(\mu)+M_p(\nu))\|u-v\|,$$

对所有 μ， $\nu \in \mathcal{W}_p(\mathbb{R}^d)$ 以及 u， $v \in \mathbb{S}^{d-1}$（定义 $M_p(\mu)$的方式请回顾公式（7）， $M_p(\nu)$）。

球面 $\mathbb{S}^{d-1}$ 是紧的，对于所有 $\varepsilon > 0$，都可以找到 $K \geq 1$ 和u_1,... , $u_K \in \mathbb{S}^{d-1}$，使得以 u_i为中心、半径为ε的球 $B(u_i, \varepsilon)$ 覆盖该球面。然后，由于 Lipschitz性质，最大切片 Wasserstein 距离可由以下量控制

$$\begin{aligned}
&\overline{SW}_p(\hat{F}_{n,X}, F_X) \\
&= \max_{u \in \mathbb{S}^{d-1}} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_*^{-1}, F_X \circ u^{-1}) \\
&\leq \max_{1 \leq k \leq K} \mathcal{W}_p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1}) + \varepsilon(M_p(\hat{F}_{n,X}) + M_p(F_X)).
\end{aligned}$$

提升至 p次方并取期望，可得

$$\begin{aligned}
&\mathbb{E}\big[\overline{SW}_p^p(\hat{F}_{n,X}, F_X)\big] \\
&\leq 3^{p-1}\mathbb{E}\big[\max_{1 \leq k \leq K} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_{k*}^{-1}, F_X \circ u_{k*}^{-1})\big] + 3^{p-1}\varepsilon^p(\mathbb{E}\big[M_p^p(\hat{F}_{n,X})\big] + \mathbb{E}\big[M_p^p(F_X)\big]).
\end{aligned}$$

第一项收敛至0 ， 由式 (17) 可知，

$$\mathbb{E}[\max_{1 \leq i \leq K} \mathcal{W}_p^p(\hat{F}_{n,X} \circ u_{i*}^{-1}, F_X \circ u_{i*}^{-1})] \longrightarrow 0.$$

第二项可由一个常数乘以 ε^p 控制，因为

$$\mathbb{E}[M_p^p(\hat{F}_{n,X})] = \mathbb{E}\big[\sum_{i=1}^n W_{ni}(X)\|Y_i\|^p\big] \leq C\mathbb{E}[\|Y\|^p]$$

（根据权重的性质 i）且

$$\mathbb{E}[M_p^p(F_X)] = \mathbb{E}\big[\mathbb{E}[\|Y\|^p \mid X]\big] = \mathbb{E}[\|Y\|^p]$$

（根据条件期望的塔性质）。令 $\varepsilon \to 0$，第二项便可任意小。由此可得 $\mathbb{E}[\overline{SW}_p^p(\hat{}F_{n,X}, F_X)] \to 0$。

$\hat{}$ 第 2 步）。由第 1 步可知， $\overline{SW}_p(F_{n,X}, F_X) \to 0$依概率收敛，或者等价地，$\hat{}F_{n,X} \to F_X$在度量空间（$\mathcal{W}_p(\mathbb{R}^d)$ $\overline{SW}_p$）中依概率收敛。定理4 说明，恒等映射从（$\mathcal{W}_p(\mathbb{R}^d)$ $\overline{SW}_p$）到（$\mathcal{W}_p(\mathbb{R}^d)$ $\mathcal{W}_p$）是连续的。由连续映射定理可知，$\hat{}F_{n,X} \to F_X$在度量空间（$\mathcal{W}_p(\mathbb{R}^d)$ $\mathcal{W}_p$）中依概率收敛。等价地， $\mathcal{W}_p(\hat{}F_{n,X}, F_X) \to 0$ 依概率收敛。

第 3 步）。由三角不等式，

$$\mathcal{W}_p(\hat{F}_{n,X}, F_X) \leq \mathcal{W}_p(\hat{F}_{n,X}, \delta_0) + \mathcal{W}_p(\delta_0, F_X)$$

其中 δ_0 是位于 0 处的狄拉克测度。此外，对于任意 $\mu \in \mathcal{W}_p(\mathbb{R}^d)$,

$$\mathcal{W}_p(\mu, \delta_0) = \left(\int_{\mathbb{R}^d} \|x\|^p \, \mu(\mathrm{d}x)\right)^{1/p} = M_p(\mu).$$

我们推得

$$\mathcal{W}_p^p(\hat{F}_{n,X}, F_X) \leq 2^{p-1} M_p^p(\hat{F}_{n,X}) + 2^{p-1} M_p^p(F_X).$$

为了证明左侧的一致可积性，只需证明

$$M_p^p(F_X) \text{ is integrable and } M_p^p(\hat{F}_{n,X}),\ n \geq 1, \text{ is uniformly integrable.} \quad (18)$$

我们有

$$M_p^p(F_X) = \mathbb{E}[\|Y\|^p \mid X]$$

它是可积的，因为 $\mathbb{E}[\|Y\|^p] < \infty$。此外，

$$M_p^p(\hat{F}_{n,X}) = \sum_{i=1}^{n} W_{ni}(X)\|Y_i\|^p$$

而 Stone 定理保证

$$\sum_{i=1}^{n} W_{ni}(X)\|Y_i\|^p \longrightarrow \mathbb{E}[\|Y\|^p \mid X] \quad \text{in } L^1.$$

由于序列 $M_p^p(\hat{F}_{n,X})$ 在 L^1 中收敛，因此它是一致可积的，结论得证。 □

4.2 命题 1、推论 1-2 和定理 3

命题 *1* 的证明。上界的证明依赖于耦合论证。不失一般性，我们可以假设 Y_i 是通过反函数法由均匀随机变量 U_i 生成的——也就是说，我们假设 U_i, $1 \leq i \leq n$ 是在 (0, 1) 上服从均匀分布、相互独立同分布，且还与协变量 X_i, $1 \leq i \leq n$ 独立的随机变量，并令 $Y_i = F_{X_i}^{-1}(U_i)$。那么，$\hat{}$ 样本 (X_i, Y_i) 是服从分布 P 的独立同分布样本。为了比较 $F_{n,x}$ 和 F_x，我们引入随机变量 $\tilde{Y}_i = F_x^{-1}(U_i)$，并定义

$$\tilde{F}_{n,x}(z) = \sum_{i=1}^{n} W_{ni}(x)\mathbb{1}_{\{\tilde{Y}_i \leq z\}}.$$

根据三角不等式，

$$\mathcal{W}_1(\hat{F}_{n,x}, F_x) \leq \mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x}) + \mathcal{W}_1(\tilde{F}_{n,x}, F_x).$$

在右侧，第一项可解释为比较加权样本 $(Y_i, W_{ni}(x))$ 与 $(Y_i, W_{ni}(x))$ 的近似˜误差，其中˜Y_i 具有目标分布 F_x。第二项是估计误差

˜这里我们使用具有正确分布的加权样本 $(Y_i, W_{ni}(x))$ 来估计 F_x。

我们首先考虑近似误差。类似于方程（16）的证明可得

$$\mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x}) \leq \sum_{i=1}^{n} W_{ni}(x)|Y_i - \tilde{Y}_i|.$$

引入均匀随机变量 U_i's，可得

$$\begin{aligned}
\mathbb{E}[\mathcal{W}_1(\hat{F}_{n,x}, \tilde{F}_{n,x})] &\leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)|F_{X_i}^{-1}(U_i) - F_x^{-1}(U_i)|\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)\int_0^1 |F_{X_i}^{-1}(u) - F_x^{-1}(u)|\,\mathrm{d}u\Big] \quad \text{by independence} \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(x)\mathcal{W}_1(F_{X_i}, F_x)\Big],
\end{aligned}$$

其中等号依赖于方程（9）。需要注意的是，这种对近似误差的控制非常一般，还可推广至阶数为 $p > 1$ 的 Wasserstein 距离。

接下来考虑估计误差，我们的方法仅适用于 $p = 1$。根据方程（10），

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,x}, F_x)] = \mathbb{E}\Big[\int_{\mathbb{R}} \Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|\mathrm{d}z\Big].$$

应用 Fubini定理并利用上界

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|\Big] \\
&\leq \mathbb{E}\Big[\Big|\sum_{i=1}^{n} W_{ni}(x)\big(\mathbb{1}_{\{\tilde{Y}_i \leq z\}} - F_x(z)\big)\Big|^2\Big]^{1/2} \\
&= \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2}\sqrt{F_x(z)(1 - F_x(z))},
\end{aligned}$$

我们推得

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,x}, F_x)] \leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(x)\Big]^{1/2}\int_{\mathbb{R}} \sqrt{F_x(z)(1 - F_x(z))}\mathrm{d}z.$$

合并这两项即可得到命题 1。 □

推论 *1* 的证明。对于采用均匀核和权重的核算法（5），记

$$N_n(X) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B(X,h_n)\}}$$

球 $B(X,h_n)$ 中的点数，其中心为 X，半径为 h_n。如果$N_n \geq 1$，则只有 $B(X,h_n)$ 中的点具有非零权重，且该权重等于$1/N_n$。如果 $N_n = 0$，则按照约定，所有权重都等于 $1/n$。因此我们推得

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}^2(X)\Big] = \mathbb{E}\Big[\frac{1}{N_n(X)}\mathbb{1}_{\{N_n(X)\geq 1\}}\Big] + \frac{1}{n}\mathbb{P}(N_n(X)=0)$$

and

$$\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(X)\|X_i - X\|^H\Big] \leq h_n^H \mathbb{P}(N_n(X)\geq 1) + k^{H/2}\mathbb{P}(N_n(X)=0)$$

因为，对于权重非零的点，到 X 的距离可以在 $N_n(X) \geq 1$ 时由上界 h_n 限定，否则由 $\sqrt{k}$ 限定（注意，$\sqrt{k}$ 是 $[0,1]^k$ 的直径）。

接下来，我们利用这一事实：给定 $X = x$，$N_n(x)$ 服从参数为 n 和 $p_n(x) = \mathbb{P}(X_1 \in B(x,h_n))$ 的二项分布。这意味着

$$\mathbb{E}\Big[\frac{1}{N_n(X)}\mathbb{1}_{\{N_n(X)\geq 1\}}\Big] \leq \mathbb{E}\Big[\frac{2}{np_n(X)}\Big] \leq \frac{2c_k}{nh_n^k}$$

其中，第一个不等式由 Györfi 等人 (2002年，引理4.1）得出，第二个不等式则由 Györfi 等人 (2002年，公式5.1）得出，其中常数$c_k = k^{k/2}$ 可以取为。类似地，

$$\begin{aligned}\mathbb{P}(N_n(X)=0) &= \mathbb{E}[(1-p_n(X))^n] \leq \mathbb{E}[e^{-np_n(X)}] \\ &\leq \big(\max_{u>0} ue^{-u}\big) \times \mathbb{E}\Big[\frac{1}{np_n(X)}\Big] \\ &\leq \frac{c_k}{nh_n^k}.\end{aligned}$$

鉴于这些不同的估计，式（13）推出

$$\begin{aligned}\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] &\leq L\Big(h_n^H + k^{H/2}\frac{c_k}{nh_n^k}\Big) + M\left(\frac{(2+1/n)c_k}{nh_n^k}\right)^{1/2} \\ &\leq Lh_n^H + M\sqrt{(2+1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.\end{aligned}$$

□

推论 *2* 的证明。对于最近邻权重（6），恰有κ_n 个非零权重，其值为 $1/\kappa_n$ 因此

$$\sum_{i=1}^{n} W_{ni}^2(X) = \frac{1}{\kappa_n}.$$

此外， κ_n 的最近邻满足

$$\|X_{i:n}(X) - X\| \leq \|X_{\kappa_n:n}(X) - X\|, \quad i = 1, \ldots, \kappa_n.$$

鉴于此，等式（13）可得

$$\begin{aligned}\mathbb{E}\big[\mathcal{W}_1(\hat{F}_{n,X}, F_X)\big] &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^H\big] + M\kappa_n^{-1/2} \\ &\leq L\mathbb{E}\big[\|X_{\kappa_n:n}(X) - X\|^2\big]^{H/2} + M\kappa_n^{-1/2}\end{aligned}$$

最后一行依据詹森不等式。根据Biau andDevroye (2015年, 定理2.4)的结论:

$$\mathbb{E}\left[\|X_{\kappa_n:n}(X) - X\|^2\right] \leq \begin{cases} 8(\kappa_n/n) & \text{if } k = 1, \\ \tilde{c}_k(\kappa_n/n)^{2/k} & \text{if } k \geq 2. \end{cases}$$

□

定理*3* （下界）的证明。Wasserstein 距离下极大极小风险下界的证明改编自Pic等人(2022年，附录C中的命题3证明；此处仅给出主要思路。

考虑 $\mathcal{D}(H, L, M)$ 的子类，其中 Y是一个可能取值为0和 B的二元变量。注意，如果1 ，定义 $B \leq 4M$的条件c）自动满足。 Y在给定 $X = x$ 时的条件分布由下式刻画:

$$p(x) = \mathbb{P}(Y = B \mid X = x)$$

而 Wasserstein 距离为

$$\mathcal{W}_1(F_x, F_{x'}) = B|p(x) - p(x')|,$$

因此，定义1 中的性质 b）等价于

$$B|p(x) - p(x')| \leq L\|x - x'\|^H. \tag{19}$$

与Pic等人（2022年，引理1）类似，可以证明，取值于 $\mathbb{R}$ 的一般预测总能改进为（就 Wasserstein误差而言）取值于 $\{0, B\}$的二元预测。事实上，对于给定的ˆ预测$F_{n,\ x}$，二元预测

$$\tilde{F}_{n,x} = (1 - \tilde{p}_n(x))\delta_0 + \tilde{p}_n(x)\delta_B$$

with

$$\tilde{p}_n(x) = \frac{1}{B}\int_0^B \big(1 - \hat{F}_{n,x}(z)\big)\mathrm{d}z$$

始终满足

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,X}, F_X)] \le \mathbb{E}[\mathcal{W}_1(\hat{F}_{n,X}, F_X)].$$

这个简单的观察意味着，在考察类别 $\mathcal{D}(H, L, M)$ 限制在二元分布上的极小极大风险时，我们可以只考虑二元预测。但对于二元预测，

$$\mathbb{E}[\mathcal{W}_1(\tilde{F}_{n,X}, F_X)] = B|\tilde{p}_n(X) - p(X)|,$$

这表明，分布回归在 Wasserstein 距离下的极小极大收敛速率，等于在正则性假设下以绝对误差估计回归函数 $\mathbb{E}[Y|X = x] = Bp(x)$ 的极小极大收敛速率（19）。根据 Stone (1980, 1982)， L^1-范数下极小极大风险的下界为 $n^{-H/(2H+k)}$ （在第一篇论文中，我们考虑了称为模型 1 示例 5 的伯努利回归模型，以及 L^q 距离，其中 $q = 1$）。

□

定理 *3* 的证明（上界）。对于核方法，推论 1指出，期望 Wasserstein 误差的上界为

$$Lh_n^H + M\sqrt{(2 + 1/n)c_k}(nh_n^k)^{-1/2} + Lk^{H/2}c_k(nh_n^k)^{-1}.$$

对右侧前两项的和关于 h_n 进行最小化，得到 $h_n \propto n^{1/(2H+1)}$ ，并意味着右侧的阶为 $n^{-H/(2H+k)}$ （最后一项可以忽略）。这与前文所述的极小极大下界收敛速率相符，并证明最优极小极大风险的阶为 $n^{-H/(2H+k)}$。

对于最近邻方法，最小化推论中期望 Wasserstein 误差的上界2 可得

$$\kappa_n \propto \begin{cases} n^{H/(H+1)} & \text{if } k = 1 \\ n^{H/(H+k/2)} & \text{if } k \ge 2 \end{cases},$$

相应的风险阶为

$$\begin{cases} n^{-H/(2H+2)} & \text{if } k = 1 \\ n^{-H/(2H+k)} & \text{if } k \ge 2 \end{cases},$$

由此可知，当 k ≥2 时，最近邻方法达到最优速率。

□

4.3 命题的证明 2

命题的证明 *2.* 第一点源于这样一个事实：与连续映射复合可以保持依概率收敛。事实上，估计量 $S(F_{n,X})$ 在 Wasserstein 距离 $\mathcal{W}_p$, $S(F_{n,X})$ 下依概率收敛到 $S(F_X)$，因此 $S(F_{n,X})$ 依概率收敛到 $S(F_X)$。

为了证明 $\mathrm{L}^{p/q}$ 中的一致性，还需要进一步证明 $|S(F_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$ 的一致可积性。利用幂函数的凸性不等式 $p/q \geq 1$，等式（14）可得

$$
\begin{aligned}
|S(\hat{F}_{n,X}) - S(F_X)|^{p/q} &\leq 2^{p/q-1}\left(|S(\hat{F}_{n,X})|^{p/q} + |S(F_X)|^{p/q}\right) \\
&\leq 2^{p/q-1}\left((aM_p^q(\hat{F}_{n,X}) + b)^{p/q} + (aM_p^q(F_X) + b)^{p/q}\right) \\
&\leq 2^{2(p/q-1)}\left(a^{p/q}M_p^p(\hat{F}_{n,X}) + a^{p/q}M_p^p(F_X) + 2b^{p/q}\right).
\end{aligned}
$$

此上界与等式（18）共同推出 $|S(F_{n,X}) - S(F_X)|^{p/q}$, $n \geq 1$ 的一致可积性，由此证明得证。 □

致谢

作者感谢法国国家研究署（ANR）通过 ANR-20-CE40-0025-01（T-REX 项目）提供支持。他们还感谢 Mehdi Dagdoug 提出了随机森林权重的示例（例 2）。

参考文献

Erhan Bayraktar 和 Gaoyue Guo。Wasserstein 类型度量之间的强等价。《电子概率通讯》，26(none):1 – 13，2021年。doi: 10.1214/21-ECP383。URLhttps://doi.org/10.1214/21-ECP383。

Gérard Biau 和 Luc Devroye。《最近邻方法讲义》。Springer 数据科学丛书。Springer，2015年。

Sergey Bobkov 和 Michel Ledoux。一维经验测度、次序统计量与 Kantorovich 输运距离。《美国数学会回忆录》，261(1259):v+126，2019年。ISSN 0065-9266。doi: 10.1090/memo/1259。URL https://doi.org/10.1090/memo/1259。

Leo Breiman。随机森林。《*Machine Learning*》，45，2001年。doi: 10.1023/a: 1010933404324。

Tilmann Gneiting 和 Matthiass Katzfuss。概率预测。《统计学及其应用年度评论》，2014。doi：10.1146/ annurev-statistics-062713-085831。

Tilmann Gneiting、Adrian E. Raftery、Anton H. Westveld 和 Tom Goldman。使用集合模型输出统计和最小 CRPS 估计进行校准概率预测。《月刊天气评论》，133 (5):1098 – 1118，2005。doi：10.1175/MWR2904.1。

J. Arthur Greenwood、J. Maciunas Landwehr、N. C. Matalas 和 J. R. Wallis。概率加权矩：定义及其与若干可用逆形式表示的分布参数之间的关系。《水资源研究》，15(5)：1049–1054，1979年。doi：https://doi.org/10.1029/WR015i005p01049。网址 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR015i005p01049。

László Györfi、Michael Kohler、Adam Krzyzak 和 Harro Walk。非参数回归的分布无关理论。施普林格统计学丛书。Springer，2002年。

Thomas M. Hamill 和 Stephen J. Colucci。eta–rsm 短期集合预报的检验。《月刊天气评论》，125，1997年6月。doi：10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2。

Alexander Henzi、Johanna F. Ziegel 和 Tilmann Gneiting。保序分布回归。《英国皇家统计学会杂志 *B* 辑：统计方法学》，83(5)：963–993，2021年。ISSN 1369-7412。

Rui Li、Brian J. Reich 和 Howard D. Bondell。深度分布回归。《计算统计与数据分析》，159：107203，2021年。ISSN 0167-9473。doi：https://doi.org/10.1016/j.csda.2021.107203。

E. A. Nadaraya。关于回归估计。《概率论及其应用》，9(1)：141–142，1964年。doi：10.1137/1109020。网址https://doi.org/10.1137/1109020。

Victor M. Panaretos 和 Yoav Zemel。水 *Wasserstein* 空间统计学入门。概率与数理统计 Springer 短论丛书。Springer，沙姆，2020年。ISBN 978-3-030-38437-1；978-3-030-38438-8。doi：10.1007/978-3-030-38438-8。网址https://doi.org/10.1007/978-3-030-38438-8。

Romain Pic、Clément Dombry、Philippe Naveau 和 Maxime Taillardat。《分布回归及其连续秩概率得分评估：极小极大风险的界与收敛》。国际预测期刊，2022年。ISSN 0169-2070。doi：https://doi.org/10.1016/j.ijforecast.2022.11.001。URLhttps://www.sciencedirect.com/science/article/pii/S0169207022001443.

Benedikt Schulz 和 Sebastian Lerch。用于阵风集合预报后处理的机器学习方法：系统比较，2021年。arXiv:2106.09512。

Erwan Scornet。随机森林的渐近性质。《多元分析杂志》，146:72–83，2016年。ISSN 0047-259X。doi: 10.1016/j.jmva.2015.06.009。URLhttps://doi.org/10.1016/j.jmva.2015.06.009。

Charles J. Stone. 一致非参数回归。《统计年鉴》，5(4):595–645，1977年。ISSN 0090-5364。网址http://links.jstor.org/sici?sici=0090-5364(197707)5:4<595:CNR>2.0.CO;2-O&origin=MSN

附有讨论以及作者的答复。

Charles J. Stone. 非参数估计量的最优收敛速率。《统计年鉴》，8(6):1348–1360，1980年。ISSN 0090-5364。网址http://links.jstor.org/sici?sici=0090-5364(198011)8:6<1348:OROCFN>2.0.CO;2-Q&origin

Charles J. Stone. 非参数回归的最优全局收敛速率。《统计年鉴》，10(4):1040–1053，1982年。ISSN0090-5364。网址http://links.jstor.org/sici?sici=0090-5364(198212)10:4<1040:OGROCF>2.0.CO;2-2&origi

Maxime Taillardat、Anne-Laure Fougères、Philippe Naveau 和 Olivier Mestre。用于降雨集合预报后处理的基于随机森林和半参数方法。《天气与预报》，34，2019年6月。doi：10.1175/WAF-D-18-0149.1。

Cédric Villani。最优传输，数学科学基础丛书第338卷/数学科学基本原理/。施普林格出版社，柏林，2009。ISBN 978-3-540-71049-3。doi: 10.1007/978-3-540-71050-9。网址https://doi.org/10.1007/978-3-540-71050-9。旧与新。

Geoffrey S. Watson。平滑回归分析。*Sankhy¯a:*《印度统计学杂志》*A*辑（*1961—2002*年），26(4):359–372，1964。ISSN 0581572X。网址http://www.jstor.org/stable/25049340。Domagoj Ćevid、Loris Michel、Jeffrey Näf、Nicolai Meinshausen 和 Peter Bühlmann。分布式随机森林：异质性调整与多元分布回归。《机器学习研究杂志》，23(333):1–79，2022年。网址http://jmlr.org/papers/v23/21-0585.html。