EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Fachbereich Mathematik

Prof. Dr. Andreas Prohl
Dr. Abhishek Chaudhary

Statistical Learning 2

Summer-Semester 2023 **Tübingen, 28.06.2023**

Homework 7

Problem 1. a) Choose $M \in \mathbb{N}_0$, and let $\mathcal{G}_M$ be the set of all polynoms on $\mathbb{R}^d$ of degree M. Let $\mathscr{P}_n \equiv \mathscr{P}_n(D_n)$ be a data-dependent partition and set

$$\mathcal{G}_M \circ \mathscr{P}_n = \Big\{ f : \mathbb{R}^d \to \mathbb{R};\ f = \sum_{\mathscr{A} \in \mathscr{P}_n} g_{\mathscr{A}} \mathbb{1}_{\mathscr{A}} \ \text{for some } g_{\mathscr{A}} \in \mathcal{G}_M \quad \forall\, \mathscr{A} \in \mathscr{P}_n \Big\}.$$

Now define accordingly (as in the lecture) the related (truncated) LS estimator

$$m_n(\mathbf{x}) = T_{\beta_n}\big(\widetilde{m}_n(\mathbf{x})\big) \qquad \forall\, \mathbf{x} \in \mathbb{R}^d.$$

Show that the same assumptions ($n \uparrow \infty$)

$$(\mathbf{a}) \qquad \beta_n \uparrow \infty\,, \quad \frac{M(\Pi_n)\cdot \beta_n^4 \cdot \log(\beta_n)}{n}\,, \quad \frac{\log\big(\Delta_n(\Pi_n)\big)\cdot \beta_n^4}{n} \downarrow 0\,, \quad \frac{\beta_n^4}{n^{1-\delta}} \downarrow 0 \qquad \text{for some } \delta > 0\,,$$

$$(\mathbf{b}) \qquad \inf_{\mathscr{S} \in \mathbb{R}^d:\, \mu[\mathscr{S}] \geq 1-\delta} \mu\Big[\Big\{ \mathrm{diam}\big(\mathscr{A}_n(\mathbf{x}) \cap \mathscr{S}\big) > \gamma \Big\}\Big] \downarrow 0 \quad \mathbb{P}\text{-f.s.} \qquad \text{for all } \gamma > 0 \text{ and } \delta \in (0,1)$$

imply strong universal consistency of this LS estimator.

b) Let $d = 1$, and $M = 1$. Use part **a)** to define a *strongly consistent* LS estimator based on data dependent partitions with statistically equivalent blocks.

Problem 2. In the lecture, we verified when *strong consistency* holds for the data dependent partitioning estimator based on 'statistically equivalent blocks/cells' — when $\mathbf{X}$ takes values in $\mathbb{R}^1$.

As is written in the book by [Györfi, p. 245], '...the concept of statistically equivalent blocks can be extended to $\mathbb{R}^d$ as follows (the so-called Gessaman rule): For fixed sample size n set $M = \lfloor (\frac{n}{k_n})^{\frac{1}{d}} \rfloor$. According to the first coordinate axis, partition the data into M sets such that the first coordinates form statistically equivalent blocks. We obtain M cylindrical sets. In the same fashion, cut each of these cylindrical sets along the second axis into M statistically equivalent blocks. Continuing in the same way along the remaining coordinate axes, we obtain M^d rectangular cells, each of which (with the exception of those on the boundary) contains k_n points (see Figure 4.6)...'

Find conditions on β_n and k_n such that the truncated data-dependent partitioning estimate, which uses a partition defined by Gessaman's rule, is strongly consistent for all distributions of $(\mathbf{X}, Y)$ where each component of $\mathbf{X}$ *has a density* and $\mathbb{E}[Y^2] < \infty$.

Problem 3. In the lecture, we tried other possible partitioning rules for data D_n — which again are based on the concept of *statistically equivalent blocks* — for situations where $\mathbf{X} = (X^1, \ldots, X^d)^\top$

takes values in $\mathbb{R}^d$, for $d \geq 2$. One strategy — seemingly efficient to fastly achieve such a partitioning $\mathscr{P}_n(\boldsymbol{z}_n)$ — recursively cuts a macro-cell into smaller ones, and the first step of it is as follows:

a) start the procedure with a macro-cell $\mathcal{R}_0$ that contains all $\boldsymbol{x}_n = \{\mathbf{x}_j\}_{j=1}^n$ — the first components in $\boldsymbol{z}_n$.

b) Identity the coordinate $\ell_0^* \in \{1, \dots d\}$ for which the *standard deviation* of $\{x_j^{\ell^*}\}_{j=1}^n$ is largest, and compute its *median value* $m_{\ell^*}(\mathcal{R}_0) \equiv m_{\ell^*}\big(\mathcal{R}_0; \{x_j^{\ell_*}\}_{j=1}^n\big)$.

c) Now decompose $\mathcal{R}_0$ into $\bigcup_{i=1}^2 \mathcal{R}_{0i}$, by locating

   (c$_1$) all $\mathbf{x}^j$ (and so also $\mathbf{z}_j \subset \boldsymbol{z}_n$) into $\mathcal{R}_{01}$ which satisfy $x_{\ell^*}^j < m_{\ell^*}(\mathcal{R}_0)$, and

   (c$_2$) into $\mathcal{R}_{02}$ the remaining data points.

The recursive construction now proceeds accordingly with the two 'macro-cells' $\mathcal{R}_{01}$ and $\mathcal{R}_{0,2}$, and comes to a stop when all cells of this partition $\mathcal{P}_n(\boldsymbol{z}_n)$ contain the (almost) same amount of data points.

Is this data-dependent partitioning rule *strongly consistent*?

Date of Submission: 12.00 on 05.07.2023.