FROM NONPARAMETRICS TO GEOMETRY: DENSITY SUPPORT ESTIMATION

Contents

1. Problem at Stake and Methodology

We observe a sample $X_1, \ldots, X_n \sim_{i.i.d.} P$ in $\mathbb{R}^d$, and we are interested in estimating the *support* $S \subset \mathbb{R}^d$ of P, that is, the smallest closed set that contains all the mass of P,

$$S = \operatorname{supp} P = \bigcap_{\substack{P(\overline{C})=1 \\ C \subset \mathbb{R}^d}} \overline{C}.$$

Throughout this chapter, we will always assume that S is compact. Assume that P is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$, and denote by $f : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$ its density. Under suitable assumptions on f — which is only defined up to a λ-negligible set —, estimating S will boil down to estimating the support of f, defined by

$$\operatorname{supp} f = \overline{\{x \in \mathbb{R}^d | f(x) > 0\}},$$

which is why this problem is often called *density support estimation.*

Proposition 1.1. *If a version $f = \mathrm{d}P/\mathrm{d}\lambda$ of the density of P is continuous on its support* $\operatorname{supp} f$*, then* $\operatorname{supp} P = \operatorname{supp} f$*.*

Proof. As $(\operatorname{supp} P)^c$ contains no mass and is open, we have $(\operatorname{supp} P)^c = \left\{x \in \mathbb{R}^d | \exists \varepsilon > 0, P(\mathrm{B}(x, \varepsilon)) = 0\right\}$. Hence,

$$\begin{aligned} \operatorname{supp} P &= \left\{x \in \mathbb{R}^d | \forall \varepsilon > 0, P(\mathrm{B}(x, \varepsilon)) > 0\right\} \\ &= \left\{x \in \mathbb{R}^d | \forall \varepsilon > 0, \int_{\mathrm{B}(x,\varepsilon)} f \mathrm{d}\lambda > 0\right\}. \end{aligned}$$

As a result, if $x \in \operatorname{supp} P$, then for all $\varepsilon > 0$, there exists $x_\varepsilon \in \mathrm{B}(x, \varepsilon)$ such that $f(x_\varepsilon) > 0$ and in particular, $x = \lim_{\varepsilon \to 0} x_\varepsilon \in \operatorname{supp} f$.

Conversely, any $x \in \operatorname{supp} f$ writes as a limit $x = \lim_{\varepsilon \to 0} x_\varepsilon$ of points $x_\varepsilon \in \mathbb{R}^d$ such that $f(x_\varepsilon) > 0$. But for all $\varepsilon > 0$, by continuity of f at $x_\varepsilon \in \operatorname{supp} f$, $\int_{\mathrm{B}(x_\varepsilon, \delta)} f \mathrm{d}\lambda > 0$ for all $\delta > 0$, so that $x_\varepsilon \in \operatorname{supp} P$. By closedness of $\operatorname{supp} P$, we get $x \in \operatorname{supp} P$. $\square$

Throughout this chapter, we will always assume that $S = \operatorname{supp} P$ is compact.

1.1. **A Direct Plugin.** A first idea could be to estimate S by the plugin $\hat{S}^0 = \overline{\{\hat{f}_n > 0\}}$, where $\hat{f}_n$ is a kernel density estimator,

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

$h = h_n$ is a properly chosen sequence of bandwidths, and $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function. The estimator $\hat{S}^0$ is a very simple and natural choice, but it presents a major limitation. Indeed, observe that $\hat{S}^0$ is compact if and only if $\operatorname{supp} K$ is compact. Hence, we are restricted to using compact-supported kernels K. In the worst case scenario, such as for the Gaussian kernel $K(x) = \exp(-\|x\|^2/2)/(2\pi)^{d/2}$, $\operatorname{supp} K = \mathbb{R}^d$, so that $\hat{S}^0$ is always $\mathbb{R}^d$.

Remark 1.2. If $\operatorname{supp} K$ is bounded and $K \geqslant 0$, the estimator $\hat{S}^0 = \{\hat{f} > 0\}$ is a finite union of rescaled translations of $\operatorname{supp} K$. That is,

$$\hat{S}^0 = \bigcup_{i=1}^{n} \operatorname{supp} K\left((\cdot - X_i)/h\right) = \bigcup_{i=1}^{n} X_i + h \operatorname{supp} K.$$

When $\operatorname{supp} K = \mathrm{B}(0,1)$, this estimator is known as the *Devroye-Wise* estimator.

1.2. **Free Thresholding.** To overcome the above limitation, we will consider a modified version of $\hat{S}^0$ by introducing a threshold parameter, in addition to the bandwidth parameter h of $\hat{f}$. Namely, we will estimate S with

$$\hat{S} = \hat{S}(f_n, \alpha_n) = \{f_n > \alpha_n\},$$

where f_n is an estimator of the density f (usually, but not necessarily, of kernel type: in this case we will denote it by $\hat{f}_n$ instead of f_n) and α_n is a sequence converging to zero.

Remark 1.3. – In contrast to its target $\operatorname{supp} f = \overline{\{x \in \mathbb{R}^d | f(x) > 0\}}$, note that the chosen estimator $\hat{S} = \{f_n > \alpha_n\}$ has no reason to be closed. Even $\hat{S} = \{f_n \geqslant \alpha_n\}$ might not be closed, since K is not assumed to be continuous: for instance, the classical rectangular kernel $K(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}$ yields discontinuous $\hat{f}_n$. All the results below would also hold for the estimators $\overline{\{f_n > \alpha_n\}}$ and $\overline{\{f_n \geqslant \alpha_n\}}$, but with extra technicalities in the proofs and without any substantial benefit. We chose to omit this feature and keep the simpler estimator $\hat{S} = \{f_n > \alpha_n\}$.

– When $K = c_d \mathbb{1}_{\mathrm{B}(0,1)}$, one easily sees that $\hat{S}^0 = \{\hat{f} > 0\} = \{\hat{f} \geqslant 1/n\}$, so that $\hat{S}(\hat{f}_n, \alpha_n)$ is a generalization of $\hat{S}^0$.

2. A L^1 Loss for Set Estimation

As the parameter of interest S is a subset of $\mathbb{R}^d$, we first need to define the notion of proximity to analyze the performance of the estimates. In other words, we shall formalize what "$\hat{S}$ is close to S" means. A standard choice comes through the Lebesgue measure-based loss defined below. Throughout this chapter, λ will denote the Lebesgue measure on $\mathbb{R}^d$.

Definition 2.1 (L^1 Distance)**.** Given two measurable sets $A, B \subset \mathbb{R}^d$, the L^1 distance between them is defined by

$$\mathrm{d}_\lambda(A, B) = \|\mathbb{1}_A - \mathbb{1}_B\|_{L^1(\mathrm{d}\lambda)}\,,$$

where $\mathbb{1}_A$ and $\mathbb{1}_B$ stand for the indicator functions of A and B.

Remark 2.2. – As a direct consequence of the definition, d_λ is a pseudo-distance: it is symmetric, satisfies the triangle inequality, and $\mathrm{d}_\lambda(A, B) = 0$ if and only if A and B differ by a Lebesgue-negligible set.
– The preceding definition uses the functional representation of sets given by $K \mapsto \mathbb{1}_K$ to provide a distance between sets.

A more geometric formulation of d_λ stands as follows

Proposition 2.3 (Measure of the Symmetric Difference). *For all measurable sets $A, B \subset \mathbb{R}^d$,*

$$\mathrm{d}_\lambda(A, B) = \lambda\,(A \triangle B)\,,$$

where $A \triangle B = (A \cap B^c) \cup (B \cap A^c) = (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference of A and B.

Proof of Proposition 2.3. Follows from the identity $|\mathbb{1}_A - \mathbb{1}_B| = \mathbb{1}_{A \triangle B}$. □

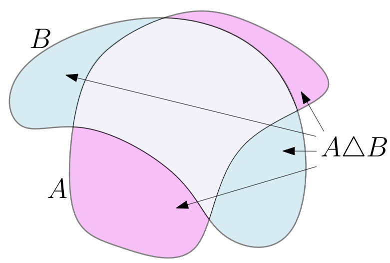

Figure 1. The symmetric difference $A \triangle B$ between two subsets A and B of the plane. Its surface corresponds to $\mathrm{d}_\lambda(A, B)$.

Remark 2.4. – The above proposition explains why d_λ is often called *measure of the symmetric difference.*
– One could take any Borel measure μ and define a pseudo-distance d_μ accordingly. It would have the same properties as d_λ. In this introductory chapter, we chose to focus on the Lebesgue measure for simplicity.

3. A Universal Consistence Result

We first prove a theorem which provides a result on consistency for the estimator (1.2) where f_n is a general density estimate.

Theorem 3.1 (Cuevas, Fraiman). *Let f be a density on $\mathbb{R}^d$ with a compact support S. Given a sequence $(f_n)_{n\geqslant 1}$ of density estimators, define an associated sequence of support estimators $\hat{S} = \{f_n > \alpha_n\}$, where $\alpha_n \searrow 0$. Assume that*

(i) $\lambda(E_0) = 0$, where $E_0 = \{x \in S | f(x) = 0\}$;
(ii) $\alpha_n^{-1} \int |f_n - f| \mathrm{d}\lambda \xrightarrow[n\to\infty]{} 0$ a.s. (resp. in probability).

Then, $\mathrm{d}_\lambda(S, \hat{S}) \xrightarrow[n\to\infty]{} 0$ a.s (resp. in probability).

Remark 3.2 (On Theorem 3.1)**.** – Condition (i) excludes pathological cases where the set $\{f > 0\}$ is far away from the support S. For instance, there exist open sets $A \subset [0, 1]$ dense in $[0, 1]$ such that $0 < \lambda(A) < 1$, such as the complement in $[0, 1]$ of a Cantor-type set of positive measure. Let f be the uniform density constant on A and null on A^c. The support of f is $[0, 1]$ and $\lambda(E_0) = 1 - \lambda(A) > 0$.
– Condition (ii) formalizes the fact that plugged in estimators f_n should converge fast enough compared to the threshold sequence α_n.

Proof of Theorem 3.1. Define $A_n = \{x \in \mathbb{R}^d || f_n(x) - f(x)| > \alpha_n\}$. Decomposing $\hat{S} \triangle S$ with respect to A_n and taking into account $\lambda(\hat{S} \cap S^c \cap A_n^c) = 0$ and $\hat{S}^c \cap S \cap A_n^c \subset \{f \leqslant 2\alpha_n\} \cap S$, we get

$$\begin{aligned}
\mathrm{d}_\lambda(S, \hat{S}) &= \lambda\left((\hat{S} \triangle S) \cap A_n\right) + \lambda\left((\hat{S} \triangle S) \cap A_n^c\right) \\
&\leqslant \lambda(A_n) + \lambda(S \cap \hat{S}^c \cap A_n^c) + \lambda(\hat{S} \cap S^c \cap A_n^c) \\
&\leqslant \lambda(A_n) + \lambda(\{f \leqslant 2\alpha_n\} \cap S).
\end{aligned}$$

From (i), $\lambda(\{f \leqslant 2\alpha_n\} \cap S) \searrow 0$ by monotone convergence, since $\{f \leqslant 2\alpha_n\} \cap S \searrow E_0$. Furthermore, from Markov inequality,

$$\lambda(A_n) = \lambda(\{|f_n - f| > \alpha_n\}) \leqslant \alpha_n^{-1} \int |f_n - f| \mathrm{d}\lambda,$$

so that $\lambda(A_n) \xrightarrow[n\to\infty]{} 0$ a.s. (resp. in probability) from (ii), which concludes the proof. $\square$

Remark 3.3. – In the case where $f_n = \hat{f}_n$ is a sequence of d-variate kernel estimators, assumption (ii) would typically be fulfilled (in probability) by a sequence α_n of type $\alpha_n^{-1} = o(n^{\frac{2k}{2k+d}})$ if f is of class $\mathcal{C}^k$.
– The sequence $a_n = \lambda(\{f < 2\alpha_n\} \cap S)$ depends directly on the way in which f "decreases to the ground". In the sharp cases where f is bounded away from zero on its support, we have $a_n = 0$ eventually. This is the most favorable situation. In general, the slower a_n decreases to zero, the worse the convergence rate f_n one can get. This is fairly intuitive, since a slow decrease of an is associated with the existence of wide "empty" areas of low probability, where f is very small, which will be underrepresented in the sample.

4. Convergence Rates Under Shape Restrictions

We will establish here a rate of convergence, on average, for the estimation of the support S. It holds in the case where the auxiliary density estimate f_n is of kernel type, under some shape restrictions on the support S.

4.1. **Distance Function and Offset.**

Let us fix a couple pieces of notation to be used in the sequel.

Definition 4.1 (Distance Function)**.** For a set $K \subset \mathbb{R}^d$, the *distance function* to K, denoted by d_K, is defined by defined by

$$\mathrm{d}_K : x \in \mathbb{R}^d \mapsto \min_{p \in K} \|x - p\| .$$

Remark 4.2. Since $\{x \in \mathbb{R}^d | \mathrm{d}_K(x) = 0\} = \overline{K}$, it is clear that d_K fully characterizes K as soon as it is closed. That is, $K \mapsto \mathrm{d}_K$ is one-to-one over the set of closed sets. Also, one easily sees that d_K is 1-Lipschitz. As a result, $K \mapsto \mathrm{d}_K$ provides a functional embedding of the set of compact subsets of $\mathbb{R}^d$. This parallels the representation $K \mapsto \mathbb{1}_K$ that we used to define d_λ (see Definition 2.1). We will use this fact in upcoming chapters to define another notion of proximity between sets: the so-called Hausdorff distance.

Definition 4.3 (Offset)**.** The *r-offset of K*, also called *tubular neighborhood* in geometry, is the set K^r of points at distance at most r of K, or equivalently the sublevel set

$$K^r := \{x \in \mathbb{R}^d | \mathrm{d}_K(x) \leqslant r\}.$$

4.2. **Covering and Packing Numbers.**

A geometric condition which will appear in a natural way has to do with the volume increase from S to S^h, as measured by the *blowing-up function*

$$\Delta(S, h) := \lambda(S^h) - \lambda(S).$$

This function provides information about the complexity of the shape S: the simpler the structure of S, the smaller $\Delta(S, h)$. Conversely, as depicted in Figure 2, the wilder $\partial S = \overline{S} \setminus \mathring{S}$, the larger $\Delta(S, h)$ can get. A typical behavior, as $h \to 0$, is $\Delta(S, h) = \mathcal{O}(h)$. As we will see later on, it is the case when the boundary ∂S is not too massive (see Lemma 4.6). To measure massiveness of ∂S, we will use packing and covering numbers. That is, roughly speaking, numbers of balls optimally displayed at some scale r in ∂S.

A r-covering of $K \subset \mathbb{R}^d$ is a subset $\mathcal{X} = \{x_1, \ldots, x_k\} \subset K$ such that for all $x \in K$, $\mathrm{d}_{\mathcal{X}}(x) \leqslant r$. A r-packing of K is a subset $\mathcal{Y} = \{y_1, \ldots, y_k\} \subset K$ such that for all $y, y' \in \mathcal{Y}$, $\mathrm{B}(y, r) \cap \mathrm{B}(y', r) = \emptyset$ (or equivalently $\|y' - y\| > 2r$).

Definition 4.4 (Covering and Packing numbers)**.** For $K \subset \mathbb{R}^d$ and $r > 0$, the covering number $\mathrm{cv}(K, r)$ is the minimum number of balls of radius r that are necessary to cover K:

$$\mathrm{cv}(K, r) = \min\{k > 0 \mid \text{there exists a } r\text{-covering of cardinality } k\}.$$

The packing number $\mathrm{pk}(K, r)$ is the maximum number of disjoint balls of radius r that can be packed in K:

$$\mathrm{pk}(K, r) = \max\{k > 0 \mid \text{there exists a } r\text{-packing of cardinality } k\}.$$

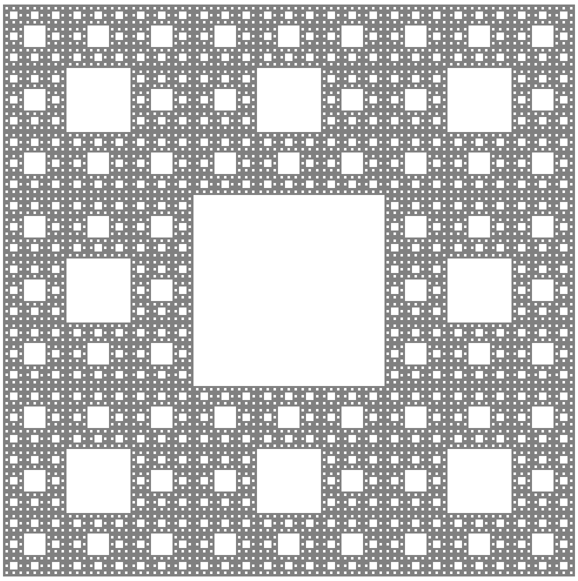

FIGURE 2. A shape S with wild boundary ∂S allows for arbitrarily large $\Delta(S,h) = \lambda(S^h \setminus S)$. Here, the so-called Sierpinski carpet.

For a given space K, covering and packing numbers usually have the same order of magnitude. Furthermore, this order of magnitude informs us about a notion of intrinsic dimension of K. Let us formalize this through two important properties of covering and packing numbers.

PROPOSITION 4.5. *Let $K \subset \mathbb{R}^d$ be a bounded subset.*

(i) For all $r > 0$,

$$\mathrm{pk}(K, 2r) \leqslant \mathrm{cv}(K, 2r) \leqslant \mathrm{pk}(K, r).$$

(ii) For all $r > 0$,

$$\mathrm{pk}(K, r) \leqslant \frac{\lambda(K^r)}{\lambda(\mathrm{B}(0, r))}.$$

In particular,

$$\mathrm{pk}(K, r) \leqslant \left(1 + \frac{\operatorname{diam} K}{r}\right)^d.$$

(iii) For all $r > 0$,

$$\mathrm{cv}(\mathrm{B}(0,1), r) \geqslant \left(\frac{1}{r}\right)^d.$$

Proof. (i) For the left-hand side inequality, notice that if K is covered by a family of balls of radius $2r$, each of these balls contains at most one point of a maximal packing $\mathcal{Y}$ at scale $2r$. Conversely, the right-hand side inequality follows from the fact that a maximal r-packing $\mathcal{Y}$ is always a $2r$-covering. If it was not the case, one could add a point x_0 such that $\mathrm{d}_{\mathcal{Y}}(x_0) > 2r$, which is impossible by maximality of $\mathcal{Y}$.

(ii) Let $\mathcal{Y} = \{y_1, \ldots, y_k\} \subset K$ be a r-packing of K. From the inclusion $\cup_{y \in \mathcal{Y}} \mathrm{B}(y, r) \subset K^r$ and the disjointness of $\mathrm{B}(y, r)$ and $\mathrm{B}(y', r)$ for all $y \neq y' \in \mathcal{Y}$, we get

$$\sum_{y \in \mathcal{Y}} \lambda(\mathrm{B}(y, r)) \leqslant \lambda(K^r),$$

which rewrites as $|\mathcal{Y}| \leqslant \lambda(K^r)/\lambda(\mathrm{B}(0, r))$ by invariance of the Lebesgue measure under translations, and yields the first claim.

For the second one, Jung's Theorem [Fed69, Theorem 2.10.41] asserts that K is contained in a (unique) closed ball with (minimal) radius at most $\sqrt{\frac{d}{2d+1}} \operatorname{diam} K$. As a result, denoting by $\omega_d = \lambda(\mathrm{B}(0, 1))$, we get

$$\frac{\lambda(K^r)}{\lambda(\mathrm{B}(0, r))} \leqslant \frac{\omega_d \left(\sqrt{\frac{d}{2d+1}} \operatorname{diam} K + r\right)^d}{\omega_d r^d} \leqslant \left(1 + \frac{\operatorname{diam} K}{r}\right)^d.$$

(iii) If $\mathcal{X} = \{x_1, \ldots, x_k\}$ is an ε-covering of $\mathrm{B}(0, 1)$, then

$$\mathrm{B}(0, 1) \subset \cup_{i=1}^k \mathrm{B}(x_i, r),$$

so

$$\lambda(\mathrm{B}(0, 1)) \leqslant k \lambda \mathrm{B}(0, r) = k r^d \lambda(\mathrm{B}(0, 1)),$$

so that $k \geqslant 1/r^d$.

□

Let us come back to the behavior of $\Delta(S, h)$ as $h \to 0$, when the boundary $\partial S = \overline{S} \setminus \mathring{S}$ of S has a controlled covering number.

Lemma 4.6. *Let $S \subset \mathbb{R}^d$ be closed. Assume that there exists $r_0 > 0$ and $C > 0$ such that for all $r \in (0, r_0)$, $\mathrm{cv}(\partial S, r) \leqslant C/r^{d-1}$. Then for all $r \in (0, r_0)$,*

$$\Delta(S, r) := \lambda(S^r) - \lambda(S) \leqslant C' r,$$

for some $C' > 0$.

Proof of Lemma 4.6. Let us first prove that $S^r \setminus S \subset (\partial S)^r$. To this aim, take $z \in S^r \setminus S$ and an associated $x \in S$ such that $\|z - x\| \leqslant r$. As the segment $[x, z]$ is connected and intersects both S and S^c, it must intersect its boundary ∂S (*lemme de passage des douanes*). Therefore, there exists $x' \in [x, z] \cap \partial S$, which means that $\mathrm{d}_{\partial S}(z) \leqslant \|z - x'\| \leqslant r$, and hence that $z \in (\partial S)^r$.

Now, let $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \partial S$ be a minimal covering of ∂S of radius r, i.e. $N = \mathrm{cv}(\partial S, r)$. From the previous point we can write

$$\begin{aligned}
\Delta(S, r) = \lambda(S^r \setminus S) \leqslant \lambda((\partial S)^r) & \\
&\leqslant \lambda\left(\left(\cup_{j=1}^N \mathrm{B}(x_i, r)\right)^r\right) \\
&= \lambda\left(\cup_{j=1}^N \mathrm{B}(x_i, 2r)\right) \\
&\leqslant \sum_{j=1}^N \lambda(\mathrm{B}(x_i, 2r)) = N \omega_d (2r)^d \leqslant 2^d C \omega_d r,
\end{aligned}$$

where $\omega_d = \lambda(\mathrm{B}(0,1))$ stands for the volume of the unit d-dimensional Euclidean ball. □

Another notion of set regularity that we will use is the *standardness*. The intuitive idea is to exclude some pathological sets, such as those having arbitrarily sharp peaks.

Definition 4.7 (Standard set)**.** A bounded set $S \subset \mathbb{R}^d$ is said to be *standard* if for every $r_0 > 0$, there exists $A \in (0,1)$ such that for all $x \in S$ and $r \in (0, r_0)$,

$$\lambda\left(S \cap \mathrm{B}(x,r)\right) \geqslant A\lambda\left(\mathrm{B}(x,r)\right) = \omega_d A r^d,$$

where $\omega_d = \lambda(\mathrm{B}(0,1))$.

Remark 4.8. – This notion is also known as the *inner cone condition* in the PDE literature.

(A) A regular set. (B) Another regular set. (C) Not a regular set.

FIGURE 3. Illustrating the notion of regularity of a set. These examples show that it prevents sets to have to sharp outwards peaks, but still allows for inwards ones.

THEOREM 4.9 ([CF97]). *Let $\hat{S} = \{\hat{f}_n > \alpha_n\}$ with $\alpha_n \to 0$, and $\hat{f}_n$ a kernel density estimator with kernel K. Assume that:*

(i) K fulfills $c_1 \mathbb{1}_{\mathrm{B}(0,r_1)} \leqslant K \leqslant c_2 \mathbb{1}_{\mathrm{B}(0,r_2)}$, for some constants $c_1, c_2 > 0$ and $0 < r_1 < r_2$, where $\mathbb{1}_A$ denotes the indicator function of the set A;
(ii) S is standard;
(iii) f is bounded away from zero on S, i.e. $S = \{f \geqslant a\}$ for some $a > 0$.

Then for n large enough,

$$\mathbb{E}\left[\mathrm{d}_\lambda(S, \hat{S})\right] \leqslant c_3 h^d \,\mathrm{cv}(S, r_1 h/2) \exp\bigl(-c_4 n h^d\bigr) + \Delta(S, r_2 h),$$

where c_3 and c_4 are positive constants. As a consequence, if we additionally assume that

(iv) $\mathrm{cv}(\partial S, r) \leqslant C/r^{d-1}$ for r small enough,

then

$$\mathbb{E}\left[\mathrm{d}_\lambda(S, \hat{S})\right] \leqslant c_5 \exp\bigl(-c_4 n h^d\bigr) + c_6 h.$$

Hence, by taking the suitable sequence $h = h_n \asymp (\log n/n)^{1/d}$*, one obtains the convergence rate*

$$\mathbb{E}\left[\mathrm{d}_\lambda(S, \hat{S})\right] \leqslant C\left(\frac{\log n}{n}\right)^{1/d}.$$

Proof of Theorem 4.9. We have

$$\mathrm{d}_\lambda(S, \hat{S}) = \lambda(\{\hat{f} > \alpha_n, f = 0\}) + \lambda(\{f > 0, \hat{f} \leqslant \alpha_n\}).$$

From assumption (i), $\{\hat{f} > \alpha_n\} \subset \{\hat{f} > 0\} \subset S^{r_2 h}$. Therefore, the first term of the right-hand side of (4.2) is easily bounded,

$$\lambda(\{\hat{f} > \alpha_n, f = 0\}) \leqslant \lambda(S^{r_2 h}) - \lambda(S) = \Delta(S, r_2 h).$$

To handle the second term of Section 4.2, let us consider a minimal covering of S with balls $B_j = \mathrm{B}(x_j, r_1 h/2)$, $x_j \in S$, $j \in \{1, \dots, N\}$ where $N = \mathrm{cv}(S, r_1 h/2)$. Then

$$\begin{aligned}\lambda(\{f > 0, \hat{f} \leqslant \alpha_n\}) = \lambda(S \cap \hat{S}^c) &\leqslant \lambda\left(\left(\cup_{j=1}^N B_j\right) \cap \hat{S}^c\right) \\ &\leqslant \sum_{j=1}^N \lambda(B_j \cap \hat{S}^c).\end{aligned}$$

Let

$$A_{n,j} = \left\{\frac{1}{nh^d}\sum_{i=1}^n \mathbb{1}_{B_j}(X_i) > \frac{\alpha_n}{c_1}\right\}.$$

Observe that the event $A_{n,j}$ is included in the event$\{B_j \subset \hat{S}\}$. To see this, assume that $A_{n,j}$ occurs and take $x \in B_j$. Then

$$\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \geqslant \frac{1}{nh^d}\sum_{i=1}^n \mathbb{1}_{\mathrm{B}(x, r_1 h)}(X_i) \geqslant \frac{c_1}{nh^d}\sum_{i=1}^n \mathbb{1}_{B_j}(X_i) > \alpha_n,$$

where the second inequality uses the fact that B_j has diameter $r_1 h$. In other words, if $A_{n,j}$ occurs, then $B_j \cap \hat{S}^c = \emptyset$, so that

$$\lambda(B_j \cap \hat{S}^c) = \lambda(B_j \cap \hat{S}^c)\mathbb{1}_{A_{n,j}^c} \leqslant \lambda(B_j)\mathbb{1}_{A_{n,j}^c}.$$

Hence, denoting $\omega_d = \lambda(\mathrm{B}(0,1))$, we have

$$\mathbb{E}\left[\sum_{j=1}^N \lambda(B_j \cap \hat{S}^c)\right] \leqslant \mathbb{E}\left[\sum_{j=1}^N \mathbb{1}_{A_{n,j}^c}\omega_d\left(\frac{r_1 h}{2}\right)^d\right] = \frac{\omega_d r_1^d}{2^d} h^d \sum_{j=1}^N \mathbb{P}(A_{n,j}^c).$$

We now need to bound the probabilities

$$\mathbb{P}(A_{n,j}^c) = \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_{B_j}(X_i) \leqslant \frac{nh^d \alpha_n}{c_1}\right)$$

from above, for $j \in \{1, \dots, N\}$. On the left-hand side, we recognize a sum of n independent and identically distributed Bernoulli variables with parameter

$$p_{n,j} := \mathbb{P}(X_i \in B_j) = \int_{B_j} f \mathrm{d}\lambda \geqslant aA\omega_d\left(\frac{r_1 h}{2}\right)^d := a_1 h^d,$$

where A is the standardness constant of S associated to $r_0 = \sup_n r_1 h_n/2$ (see Definition 4.7), and a is such that $f \geqslant a > 0$ on S. As a result,

$$p_{n,j} := \mathbb{P}(X_i \in B_j) = \int_{B_j} f \mathrm{d}\lambda \geqslant aA\omega_d \left(\frac{r_1 h}{2}\right)^d := a_1 h^d,$$

After centering the variables, we hence obtain

$$\begin{aligned}
\mathbb{P}(A_{n,j}^c) &= \mathbb{P}\left(\sum_{i=1}^n (\mathbb{1}_{B_j}(X_i) - p_{n,j}) \leqslant \frac{nh^d\alpha_n}{c_1} - np_{n,j}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n (\mathbb{1}_{B_j}(X_i) - p_{n,j}) \leqslant -\left(1 - \frac{h^d\alpha_n}{c_1 p_{n,j}}\right) np_{n,j}\right).
\end{aligned}$$

Since $\alpha_n \to 0$, we have

$$1 - \frac{h^d\alpha_n}{c_1 p_{n,j}} \geqslant 1 - \frac{\alpha_n}{c_1 a_1} \to_{n\to\infty} 1.$$

In particular, for n large enough we get that $1 - \frac{h^d\alpha_n}{c_1 p_{n,j}} \geqslant 1/2$, in which case

$$\mathbb{P}(A_{n,j}^c) \leqslant \mathbb{P}\left(\sum_{i=1}^n (\mathbb{1}_{B_j}(X_i) - p_{n,j}) \leqslant -np_{n,j}/2\right).$$

Now, the variables $Z_i = \mathbb{1}_{B_j}(X_i) - p_{n,j}$ are centered and satisfy $v := \sum_{i=1}^n \mathbb{E}[Z_i^2] = np_{n,j}(1-p_{n,j}) \leqslant np_{n,j}$ and $Z_i \geqslant -p_{n,j} =: -b$, Bernstein's concentration inequality [BLM13, Corollary 2.11 and (2.10)] applied with $t = np_{n,j}/2$ yields

$$\begin{aligned}
\mathbb{P}(A_{n,j}^c) &\leqslant \exp\left(-\frac{t^2}{2\big(v + bt/3\big)}\right) \\
&\leqslant \exp\left(-\frac{1}{2}\frac{\big(np_{n,j}/2\big)^2}{np_{n,j} + \big(np_{n,j}/2\big)p_{n,j}/3}\right) \\
&= \exp\left(-\frac{1}{2}\frac{np_{n,j}/4}{1 + p_{n,j}/6}\right) \\
&\leqslant \exp\left(-\frac{3}{28} n a_1 h^d\right),
\end{aligned}$$

where the last inequality uses $a_1 h^d \leqslant p_{n,j} \leqslant 1$. As a result, we get the first claim with $c_3 = 2\omega_d r_1^d 2^{-d}$, $c_4 = 3a_1/28$.

Using the extra assumption (iv), we get the second claim using Proposition 4.5 (i) and Lemma 4.6.

Plugin $h = c_7(\log n/n)^{1/d}$ for $c_7 \geqslant 1/(c_4 d)^{1/d}$ yields the last expected loss bound with $C = c_5 + c_6 c_7$. □

5. Further Sources

These notes mainly follow [CF97].

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[CF97] Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *Ann. Statist.*, 25(6):2300–2312, 1997.

[Fed69] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.