Mathematics of Machine Learning

Lecture 05/01/17

Outline

1 - Non-parametrics: an introduction
2 - Lower bounds
3 - Stochastic bandit models
4 - Classification: a theory
5 - Machine Learning and optimization
6 - Individual sequences

Prof: Aurélien Garivier
Sebastien ~~Gadat~~ GERCHINOVITZ

Lectures: 6

(1) Non-parametrics: an introduction

I Density Estimation

Literature: Tsybakov "~~Nonparametrical statistics~~" "Introduction to non-parametric estimation"

Model: $X_1, \dots, X_n \overset{iid}{\sim} P$, $P \in \mathcal{M}$ (family of proba. distributions)

Problem: we don't know P, we want to estimate it.

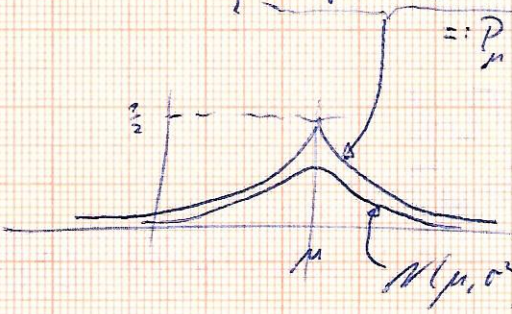
Example (parametric): $\mathcal{M} := \{ \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R} \}$, $\sigma^2 > 0$ is ~~known~~ given

→ estimate μ:

$$\hat{\mu}_n := \bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example $\mathcal{M} = \{ \frac{1}{2} \exp(-|x-\mu|) \, d\lambda, \mu \in \mathbb{R} \}$ (λ: Lebesgue measure on $\mathbb{R}$)

$=: P_\mu$



- we could estimate μ with the Maximum-likelihood-method or the moment method:
  - The expectation of P_μ is $\mathbb{E}_\mu[X_1] = \mu$
  - Law of large numbers (LLN) yields $\bar{X}_n \to \mathbb{E}_\mu[X_1] = \mu$

$$\hat{\mu}_n = \bar{X}_n$$

- for every estimator $\hat{\mu}_n$, we define the quadratic risk

$$R(\hat{\mu}_n, \mu) = \mathbb{E}_\mu\left[(\hat{\mu}_n - \mu)^2\right]$$ and one can compute

$$\begin{aligned}\mathbb{E}_\mu\left[(\hat{\mu}_n - \mu)^2\right] &= \mathbb{E}_\mu\left[(\hat{\mu}_n - \mathbb{E}_\mu[\hat{\mu}_n] + \mathbb{E}_\mu[\hat{\mu}_n] - \mu)^2\right] \\ &= \mathbb{E}_\mu\left[(\hat{\mu}_n - \mathbb{E}_\mu[\hat{\mu}_n])^2\right] \\ &\quad + (\mathbb{E}_\mu[\hat{\mu}_n] - \mu)^2 \\ &\quad + \underbrace{\mathbb{E}_\mu\left[\hat{\mu}_n - \mathbb{E}_\mu[\hat{\mu}_n]\right]}_{=0} \cdot (\mathbb{E}_\mu[\hat{\mu}_n] - \mu) \\ &= \mathrm{Var}(\hat{\mu}_n) + \underbrace{(\mathbb{E}_\mu[\hat{\mu}_n] - \mu)^2}_{=:(\mathrm{bias}(\hat{\mu}_n))^2}\end{aligned}$$

("variance-bias-decomposition")

Exp.: $R(\bar{X}_n, \mu) = \frac{\mathrm{Var}(X_1)}{n}$

Remark:
One can prove: $\exists C > 0 : \forall \hat{\mu}_n,\ \mathbb{E}_\mu\left[(\hat{\mu}_n - \mu)^2\right] \geq \frac{C}{n}$

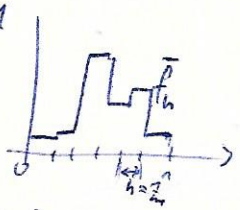
Coming to non-parametrics:

exp.: $\mathcal{M} = \left\{ dP(x) = f(x)dx,\ f \in C^2([0,1]) \right\}$

$f \in \mathcal{M}$ may look like 

(the X_i are samples that we observed)

-> How to estimate f?
-> Idea: make an histogram:



that means we introduce a (regular) partition of $[0,1]$: $C_1, \dots, C_m$ with $C_j = \left[\frac{j-1}{m}, \frac{j}{m}\right[$, $j = 1, \dots, m$

$$\bar{f}_h(x) := \sum_{j=1}^{m} \frac{p_j}{h} \mathbb{1}_{C_j}(x) \quad \text{where } h = \frac{1}{m},\ p_j := P(X_1 \in C_j)$$

$$\rightsquigarrow \frac{p_j}{h} = \frac{1}{h}\int_{C_j} f(x)dx$$ -> We estimate $\bar{f}_h$ instead of f by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} \mathbb{1}_{C_j}(X_i)\mathbb{1}_{C_j}(x)\right)$$

$$\mathcal{M}_h = \left\{ f : [0,1] \to \mathbb{R}_+,\ f(x) = \sum_{j=1}^{m} \alpha_j \mathbb{1}_{C_j}(x),\ (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m \right\}$$

$\hat{\alpha}_j = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i)$ $\quad B(p_j)$ (Bernoulli-distr. with parameter p_j)

$\mathbb{E}[\hat{\alpha}_j] = \frac{1}{h} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{C_j}(X_i)\right] = \frac{p_j}{h} = \alpha_j$

$\rightarrow \forall x \in [0,1] : \mathbb{E}[\hat{f}_h(x)] = \bar{f}_h(x)$

Risk at $x \in [0,1]$

$$MSE_f(x,h) = \mathbb{E}_f\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] = \underbrace{\left(\mathbb{E}_f[\hat{f}_h(x)] - f(x)\right)^2}_{\text{squared bias}} + \underbrace{Var_f[\hat{f}_h(x)]}_{\text{variance}}$$

$\hat{f}_h(x) = \frac{\hat{p}_{j(x)}}{h} = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{C_{j(x)}}(X_i) = \frac{Z_j}{nh}$ where $Z_j \sim B(n, p_{j(x)})$ (binomial distr.)

(j: the j for which holds that $x \in C_j$)

$\rightarrow \mathbb{E}[\hat{f}_h(x)] = \frac{p_{j(x)}}{h}$

$Var(\hat{f}_h(x)) = \frac{n\, p_{j(x)}(1 - p_{j(x)})}{(nh)^2}$

$= \frac{p_{j(x)}(1 - p_{j(x)})}{nh^2}$

$\rightarrow$ if we want to have $MSE_f(x,h) \xrightarrow{n\to\infty} 0$

we need to choose $h = h_n$ and $h_n \xrightarrow{n\to\infty} 0$

Integrated risk

$$MISE_f(h) := \int_0^1 MSE_f(x,h)\,dx \overset{\text{Fubini}}{=} \mathbb{E}_f\left[\int_0^1 \left(\hat{f}_h(x) - f(x)\right)^2 dx\right]$$

Variance: $\int_0^1 Var_f\left(\hat{f}_h(x)\right)dx = \sum_{j=1}^{m} \int_{C_j} Var_f\left[\hat{f}_h(x)\right]dx = \sum_{j=1}^{m} \frac{p_j(1-p_j)}{nh} = \frac{1}{nh} \underbrace{\sum_{j=1}^{m} p_j}_{=1} - \frac{1}{n} \sum_{j=1}^{m} p_j^2$

Bias term: $\int_0^1 \left(\mathbb{E}_f\left[\hat{f}_h(x)\right] - f(x)\right)^2 dx = \sum_{j=1}^{m} \int_{C_j} \left(\frac{p_j}{h} - f(x)\right)^2 dx$

$= \sum_{j=1}^{m} \left\{ \frac{p_j^2}{h} - 2\frac{p_j}{h} \int_{C_j} f(x)\,dx + \int_{C_j} f^2(x)\,dx \right\}$

$= \int_0^1 f^2(x)\,dx - \frac{1}{h} \sum_{j=1}^{m} p_j^2$

Lemma: If $X_1, \dots, X_n \overset{iid}{\sim}$ with density f on $[0,1]$ and if $\hat{f}_h$ is the histogram estimator with $m = \frac{1}{h}$ bins, then

$$MISE_f(h) = \mathbb{E}_f\left[\int_0^1 (\hat{f}_h(x) - f(x))^2 dx\right] = \int_0^1 f^2(x)\,dx + \frac{1}{nh} - \frac{n+1}{nh}\sum p_j^2$$

Question: How does it behave when $n \to \infty$ and $h_n \to 0$

- $$\int_{C_j} f^2(x)dx - \frac{1}{h}p_j^2 = \int_{C_j}\left(f(x) - \frac{1}{h}\int_{C_j} f(u)du\right)^2 dx = \frac{1}{h^2}\int_{C_j}\left(\int_{C_j}(f(x) - f(u))du\right)^2 dx$$

As f is assumed to be twice continuously differentiable,

$$f(u) - f(x) = (u-x)f'(a_j) + O(h^2) \quad \text{where } a_j = \frac{j-1}{m}$$

$$\int_{C_j} f^2(x)\,dx - \frac{1}{h}p_j^2 = \frac{f'(a_j)^2}{h^2}\int_{C_j}\left(\int_{C_j}(x-u)du\right)^2 dx + O(h^4)$$

Now: $$\int_{C_j}\left(\int_{C_j} x - u\,du\right)dx = h^5\int_0^1\left(\int_0^1 (y - \tfrac{1}{2})dz\right)^2 dy = \frac{h^5}{12}$$

$$\Rightarrow \int_{C_j} f^2(x)\,dx - \frac{p_j^2}{h} = \frac{h^3}{12}f'(a_j) + O(h^4)$$

$$= \frac{h^2}{12}\int_{C_j} f'(x)dx + O(h^4) \quad \text{and}$$

$$MISE_f(h) = \sum_{j=1}^{m}\left(\int_{C_j} f^2(x)dx - \frac{p_j^2}{h}\right) + \frac{1}{nh} - \frac{1}{nh}\sum p_j^2$$

$$= \frac{h^2}{12}\int_0^1 f'(x)^2 dx + O(h^3) + \frac{1}{nh} + O\left(\frac{1}{n}\right)$$

Theorem: If $X_1, \dots, X_n$ is a sample of a probab. distr. with density in $C^2([0,1])$ and if $h_n \overset{n\to\infty}{\to} 0$, then the histogram estimator $\hat{f}_{h_n}$ satisfies

$$MISE_f^{(h_n)} = \frac{h_n^2}{12}\int_0^1 f'(x)^2 dx + \frac{1}{nh_n} + O(h_n^3) + O\left(\frac{1}{n}\right) \quad \text{as } n \to \infty$$

→ in order to find the best number of bins, we need to minimize

$$g(h) = \frac{h^2}{12}\int_0^1 f'(x)^2 dx + \frac{1}{nh} \xrightarrow{\text{obtain}} h_{opt} = \left(\frac{n}{6}\int_0^1 f'(x)^2 dx\right)^{-\frac{1}{3}}$$

For that choice, $MISE_f(h_{opt}) = \left(\frac{3}{4}\right)^{2/3}\left(\int_0^1 f'(x)^2 dx\right)^{\frac{1}{3}} \cdot \frac{1}{n^{2/3}} + O\left(\frac{1}{n}\right)$.

Problems: - Can we do better ? (faster convergence) concerning the upper bound of MSE

- we don't know $f' \Rightarrow$ we don't know h_{opt}

Memo: $MISE_f(h) = \int_0^1 f^2(x)dx + \frac{1}{nh} - \frac{n+1}{nh}\sum_{j=1}^{m} p_j^2$

We want to find h minimizing $MISE_f(h)$ just using $X_1, \dots, X_n$ (and not anything unknown of f)

Remark: $\underset{h}{\operatorname{argmin}}\, MISE_f(h) = \underset{h}{\operatorname{argmin}} \left(MISE_f(h) - \int_0^1 f^2(x)dx \right)$

$= \underset{h}{\operatorname{argmin}} \left(\frac{1}{nh} - \frac{n+1}{nh}\sum_{j=1}^{m} p_j^2 \right)$

Idea: estimate $J_f(h) = MISE_f(h) - \int_0^1 f^2(x)dx$

$= \frac{1}{nh} - \frac{n+1}{nh}\sum_{j=1}^{m} p_j^2$ by

$\hat{J}_f(h) = \frac{1}{nh} - \frac{n+1}{nh}\sum_{j=1}^{m} \hat{p}_j^2$ not good because biased:

$\mathbb{E}[\hat{p}_j^2] = \mathbb{E}[\hat{p}_j]^2 + Var[\hat{p}_j] = p_j^2 + \frac{p_j(1-p_j)}{n} = p_j^2\left(1 - \frac{1}{n}\right) + \frac{p_j}{n}$

$\rightarrow \mathbb{E}\left[\hat{p}_j^2 - \frac{\hat{p}_j}{n}\right] = p_j^2\left(\frac{n-1}{n}\right) \Leftrightarrow \mathbb{E}\left[\frac{n}{n-1}\left(\hat{p}_j^2 - \frac{\hat{p}_j}{n}\right)\right] = p_j^2$

$= \frac{n}{n-1}p_j^2 - \frac{\hat{p}_j}{n-1}$

$\hat{J}_f(h) = \frac{1}{nh} + \frac{n+1}{nh}\sum_{j=1}^{m}\left(\frac{n}{n-1}\hat{p}_j^2 - \frac{\hat{p}_j}{n-1}\right)$

$= \frac{1}{nh} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{n+1}{n(n-1)h}$

$= -\frac{n+1}{(n-1)h}\sum_{j=1}^{m}\hat{p}_j^2 + \frac{1}{nh}\underbrace{\left(n + \frac{n+1}{n-1}\right)}_{=\frac{2n}{n-1}} = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\hat{p}_j^2$

Prop: $\mathbb{E}\left[\hat{J}_f(h)\right] = J_f(h)$ "cv" = "cross-validation"

Alg: Initialize: $m = 1$, $m_{cv} = 1$, $J_{cv} = +\infty$

while $m < n$ do

$J = \frac{2m}{n-1} - \frac{(n+1)m}{n-1}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{C_j}(X_i)\right)^2$

if $J < J_{cv}$ then

$m_{cv} = m$

$J_{cv} = J$

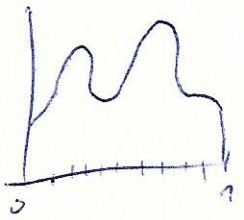
(Algo to find the best bin-width for our histogram)

endif

$m = m+1$

return $\hat{h} = \frac{1}{m_{cv}}$ → Find estimator $\hat{f}_{\hat{h}}$

Other ideas to estimate f:



coming from histograms:

if x is at the center of a bin: only

$$\hat{f}_h^{hist}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}\left\{|X_i - x| < \frac{h}{2}\right\}$$

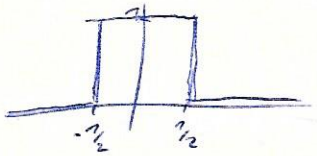
(But the center of a bin the histogram estimator coincides with the rectangular kernel estimator!)

↳ "Rectangular"

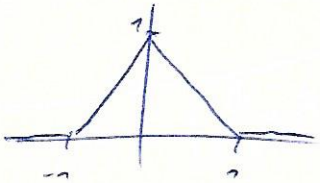
$$\hat{f}_h^{R}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}\left\{|X_i - x| < \frac{h}{2}\right\} \quad \forall x \in [0,1]$$

Def: $\hat{f}_h^{K}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$ "Kernel estimator", where K is a Kernel fct. and h a width parameter

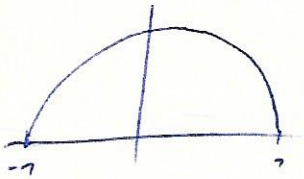
Kernel exp.: • rectangular kernel:



• triangular kernel

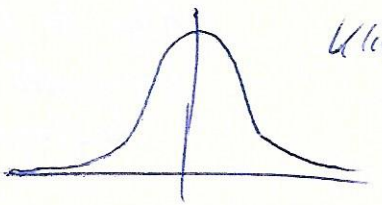


• Epanechnikov



$$K(u) = \frac{3}{4}(1-u^2)\,\mathbb{1}_{[-1,1]}(u)$$

• Gaussian kernel



$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

Exercise if $K(u) \geq 0 \ \forall u \in \mathbb{R}$ and if $\int_{-\infty}^{\infty} K(u)\,du = 1$, then f_h^K is a density function.

Remark: For now we only saw positive kernels, but there are also kernels that are partly negative (difference to densities)

What is the risk of the kernel density estimator?

Assumptions on the kernel

(1) $\int_{-\infty}^{\infty} K(u)\,du = 1$

(2) $\int_{-\infty}^{\infty} u K(u)\,du = 0 \quad (\Leftarrow K \text{ is even})$

(3) $\int_{-\infty}^{\infty} u^2 |K(u)|\,du < \infty$

(4) $\int_{-\infty}^{\infty} K(u)^2\,du < \infty$

Prop. a) Under assumptions (1), (2), (3), if f is a bounded density with a bounded second derivative, then

$$\text{bias}\left(\hat{f}_h^K(x)\right) \leq C_1 h^2 \quad \text{where} \quad C_1 = \frac{1}{2} \sup_z |f''(z)| \cdot \int_{-\infty}^{\infty} u^2 |K(u)|\,du$$

b) In addition, under conditions (1), (4)

$$\text{Var}\left(\hat{f}_h^K(x)\right) \leq \frac{C_2}{nh} \quad \text{where} \quad C_2 = \sup_z f(z) \int_{-\infty}^{\infty} K(u)^2\,du$$

Proof: a) $\mathbb{E}_f\left[\hat{f}_h^K(x)\right] = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{E}_f\left[K\left(\frac{X_i - x}{h}\right)\right] = \frac{1}{h} \mathbb{E}_f\left[K\left(\frac{X_1 - x}{h}\right)\right] = \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{y-x}{h}\right) f(y)\,dy$

$$\overset{y = x+uh}{=} \int_{-\infty}^{\infty} K(u) f(x+uh)\,du$$

Using a Taylor expansion:

$$\mathbb{E}_f\left[\hat{f}_h^K(x)\right] = \int_{-\infty}^{\infty} K(u) f(x+uh)\,du = \int_{-\infty}^{\infty} K(u)\left(f(x) - uh f'(x) + \frac{(uh)^2}{2} f''(\xi_u)\right)du$$

where $\xi_u \in [x, x+uh]$

$$= \int_{-\infty}^{\infty} K(u)\,du + h f(x)\int_{-\infty}^{\infty} uK(u)\,du + \frac{h^2}{2}\int_{-\infty}^{\infty} u^2 K(u) f''(\xi_u)\,du$$

Thus,

$$\left|\text{Bias}\left(\hat{f}_h^K(x)\right)\right| = \left|\mathbb{E}\left[\hat{f}_h^K(x)\right] - f(x)\right|$$

$$\leq \frac{h^2}{2}\left|\int u^2 K(u) f''(\xi_u)\,du\right|$$

$$\leq \frac{h^2}{2}\int u^2 |K(u)| \cdot |f''(\xi_u)|\,du$$

$$\leq h^2 \cdot \underbrace{\frac{1}{2}\sup_t |f''(t)| \cdot \int u^2 |K(u)|\,du}_{= C_1}$$

b) $$\left|\text{Var}\left(\hat{f}_h^K(x)\right)\right| = \frac{1}{(nh)^2}\text{Var}_f\left[\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right] = \frac{1}{nh^2}\text{Var}\left[K\left(\frac{X_1 - x}{h}\right)\right]$$

$$\leq \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)^2\right]$$

$$= \frac{1}{nh^2}\int_{-\infty}^{\infty} K\left(\frac{y-x}{h}\right)^2 f(y)\,dy$$

$$= \frac{1}{nh}\int_{-\infty}^{\infty} K(u)^2 f(x+uh)\,du \leq \frac{1}{nh}\sup|f|\int K(u)^2\,du$$

□

Consequence

$$MSE\left(\hat{f}_h^K(x)\right) \leq C_1^2 h^4 + \frac{C_2}{nh}$$

→ we can optimize in h to find the best bandwidth

$$h_{opt} = \left(\frac{C_2}{4C_1^2}\right)^{1/5} n^{-1/5}$$ which leads to

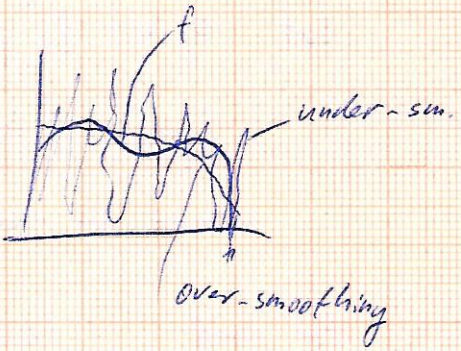
$$MSE\left(\hat{f}_{h_{opt}}^K(x)\right) \leq C n^{-4/5}$$

Remarks

- $n^{-4/5}$ is better than $n^{-2/3}$
- $n^{-4/5}$ is optimal with those assumptions on f
- if we assume that f is in $C^k(\mathbb{R}^d)$ then one can show that the optimal rate is

$$n^{\left(-\frac{2k}{2k+d}\right)} \qquad \text{(we did } k=2, d=1\text{)}$$

- h too small → "under smoothing"
- h too big → "over-smoothing"



Proposition $\hat{J}(h) = \|\hat{f}_h^k\|_2^2 - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right)$ is an unbiased estimator of $J(h) = MISE(\hat{f}_h^k) - \|f\|_2^2$

Proof:

$$\mathbb{E}_f\left[\hat{J}(h)\right] = \mathbb{E}_f\left[\|\hat{f}_h^k\|_2^2\right] - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathbb{E}_f\left[K\left(\frac{X_i - X_j}{h}\right)\right]$$

$$= \mathbb{E}_f\left[\|\hat{f}_h^k\|_2^2\right] - \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\int_{\mathbb{R}^2} K\left(\frac{x-y}{h}\right) f(x) f(y)\, dx\, dy$$

$$= \mathbb{E}_f\left[\|\hat{f}_h^k\|_2^2\right] - \frac{2}{h}\int_{\mathbb{R}^2} K\left(\frac{x-y}{h}\right) f(x) f(y)\, dx\, dy \qquad (*)$$

But $J(h) = \mathbb{E}_f\left[\|\hat{f}_h^k - f\|_2^2\right] - \|f\|_2^2$

$$= \mathbb{E}_f\left[\|\hat{f}_h^k\|_2^2\right] - 2\,\mathbb{E}_f\left[\int_{\mathbb{R}^2} \hat{f}_h^k(x) f(x)\, dx\right]$$

$$= \mathbb{E}_f\left[\|\hat{f}_h^k\|_2^2\right] - 2\int_{\mathbb{R}^2} \mathbb{E}\left[\hat{f}_h^k(x)\right] f(x)\, dx$$

$$= \mathbb{E}_f\left[\|\hat{f}_h^k\|_2^2\right] - 2\int_{\mathbb{R}^2} \frac{1}{h}\int K\left(\frac{y-x}{h}\right) f(y) f(x)\, dy\, dx$$

$$\overset{(*)}{=} \mathbb{E}_f\left[\hat{J}(h)\right] \qquad \square$$

Algo:

- compute $\hat{h} = \operatorname{argmin} \hat{J}(h)$ (for exp. on a grid of h) of possible values
- return $\hat{f}_{\hat{h}}^k$

Alternative method for chosing the bandwidth h : Lepski's method

Fix $x \in [0,1]$, $\text{supp}(f) \subseteq [0,1]$

"locally constant"

$$\hat{f}^{LC}_{n,h}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{(x-h, x+h]}(X_i)$$

Let us consider a finite sequence $h_1 < h_2 < \dots < h_M$ of possible bandwidths.

Let $e_j(x) = \mathbb{E}\left[\hat{f}_j(x)\right] - f(x)$ "approximation error" (= bias)

$\xi_j(x) = \hat{f}_j(x) - \mathbb{E}\left[\hat{f}_j(x)\right]$ "estimation error" (= variation)

where we use the notation "$\hat{f}_j = \hat{f}^{LC}_{n,h_j}$". ($n$ is fixed)

For small j, $|e_j(x)| \ll |\xi_j(x)|$.

For large j, $|e_j(x)| \gg |\xi_j(x)|$. Let $\delta > 0$.

Assume that you can choose a sequence $\mathcal{E}_1 < \mathcal{E}_2 < \dots < \mathcal{E}_M$ such that

$\mathbb{P}\left(|\xi_j(x)| \le \mathcal{E}_j, \ \forall j = 1, \dots, M \right) \ge 1 - \delta$. Then build the interval

$I_j = \left[\hat{f}_j(x) - 2\mathcal{E}_j, \ \hat{f}_j(x) + 2\mathcal{E}_j \right]$. We want to find the bandwidth h_{j^*} for which $|e_{j^*}(x)| \approx |\xi_{j^*}(x)|$.

Let $j^* = \min\{ j : |e_j| \ge \mathcal{E}_j \} - 1$

$= \max\{ j : |e_j| \le \mathcal{E}_j \}$. Then for all $j < j^*$, with high proba.

$|\hat{f}_j(x) - f(x)| \le |e_j(x)| + |\xi_j(x)| \le |e_j(x)| + \mathcal{E}_j \le 2\mathcal{E}_j$ which is equivalent to

$$f(x) \in \left[\hat{f}_j(x) - 2\mathcal{E}_j, \ \hat{f}_j(x) + 2\mathcal{E}_j \right] = I_j$$

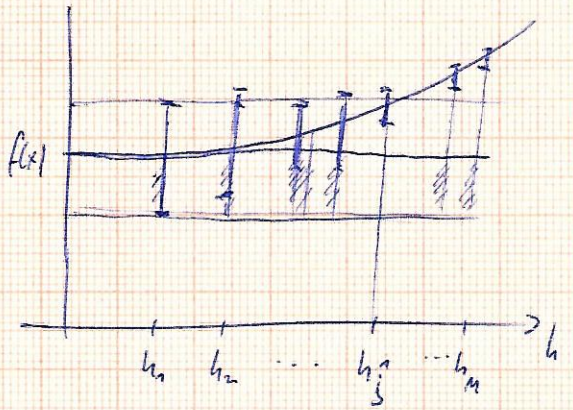
Thus, $\mathbb{P}\left(f(x) \in \bigcap_{j=1}^{j^*} I_j \right) \simeq 1 \Rightarrow \bigcap_{j=1}^{j^*} I_j \neq \emptyset$ and we define:

$$\hat{j} = \max\left\{ J = 1, \dots, M : \bigcap_{j=1}^{J} I_j \neq \emptyset \right\}$$

and $\hat{f}(x) = \hat{f}_{\hat{j}}(x)$.

Lemma: on $\Omega_0 = \left\{ \max_{1 \le j \le m} \left| \frac{\hat{f}_j(x)}{e_j} \right| \le 1 \right\}$ it holds that

$$\left| \hat{f}_{\hat{j}}(x) - f(x) \right| \le 6 \varepsilon_{j^*}$$



Assumption : $\forall x \in [0,1]$, $f(x) \le L$ and L is known.

Parameter: ε (level of confidence)

Initialization: $j = 0$, $D = [0, L]$

while $D \neq \emptyset$ and $j < m$:

$j = j + 1$, $h = h_j$

$$\hat{f} = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{[x-h, x+h]}(X_i)$$

$$\varepsilon = \sqrt{\frac{3L}{2nh} \ln\left(\frac{2m}{\varepsilon}\right)}$$

$$I = \left[\hat{f} - 2\varepsilon , \hat{f} + 2\varepsilon \right]$$

$$D = D \cap I$$

endwhile

Return: $\hat{j} = j - 1$

$\hat{h} = h_{\hat{j}}$

$\hat{f} = \hat{f}_{\hat{h}, \hat{h}}$

---

theorem: if $h_1 < h_2 < \dots < h_m$ and if f is a density on $[0,1]$ bounded by L, if $\varepsilon_j = \frac{3L}{2nh_j} \ln\left(\frac{2m}{1-\varepsilon}\right)$ then with proba. at least $1-\varepsilon$

$$j^* = \max \left\{ j \in \{1, \dots, m\} : \left| \mathbb{E}\left[\hat{f}_{h,h}(x) \right] - f(x) \right| \le \varepsilon_j \right\}$$

and if $h_j = h_1 \cdot a^{j-1}$ $(a > 1)$, if $\left| \mathbb{E}\left[\hat{f}^{LC}_{n,h}(x) \right] - f(x) \right| \leq C h^{\beta}$

then with prob. $\geq 1 - \varepsilon$

$$\left| \hat{f}_{\hat{j}}(x) - f(x) \right| \leq \text{Constant} \cdot n^{-\frac{\beta}{2\beta+1}} \cdot \left(\ln \frac{n}{\varepsilon} \right)^{\frac{\beta}{2\beta+1}}$$