# $L_1$-MDE OF A REGRESSION TYPE FUNCTION WITH RATES OF CONVERGENCE

Yannis G. Yatracos, Yau Mathematical Sciences Center, Tsinghua University

November 26, 2020

CONTENT

# 1 Motivation-The Results

It is a well-known fact that $L_r$-optimal estimates ($1 \leq r \leq \infty$) of a density and a regression function with the same smoothness converge to the true parameter at the same rate; for example:

*"... The results on optimal rates of convergence for nonparametric density estimates are surprisingly similar to those for nonparametric regression ..." (Stone (1982), Annals of Statistics, 10, p. 1044, 1. 7-8).*

The following questions arise naturally, given the presented results on density estimation:
**1.** Is there an explanation for this coincidence in the rates of convergence?
**2.** Would the same optimal rates have been observed if, other things being equal, the regression functions were a quantile or another parameter of the conditional density?

These questions provided the motivation for the regression type problem studied in these lectures. The key observation to answer both questions is that *a regression type problem can be viewed as a combination of several density estimation problems,* each occurring at the observed values of the independent variable; see graph.

In the classical regression problem $(X_1, Y_1), \ldots, (X_n, Y_n)$ are observed, according to the model

$$Y = \theta(X) + \epsilon, \tag{1}$$

$\theta$ belongs to a space of functions $\Theta$ and the assumptions for the error $\epsilon$ are:

$$E(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2, \quad \epsilon \sim P_{0,\sigma^2} \text{ unknown}$$

AIM: Estimate $\theta(x)$ (conditional mean).

Observe that: $Y_i | X_i = x_i \sim P_{\theta(x_i),\sigma^2}, \quad \theta(x_i) = E(Y_i | X_i = x_i)$.

In a regression type problem it is assumed instead of (1) that

$$Y | X = x \sim P_{\theta(x)}, \tag{2}$$

*i.e.* the regression type function $\theta(x)$ can be any parameter of the conditional probability measure $P_{\theta(x)}$, *i.e.* a conditional median or another conditional quantile.

What are the results, briefly: Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample, denote the conditional density of $Y_i|X_i = x_i$ as $f(y|x_i, \theta(x_i)$ and $\theta$ an element of a metric space, $(\Theta, d)$, of functions. A lower bound is provided for the $d$-error in estimating $\theta$. The order of the bound depends on the local behavior of the Kullback-Leibler information of the conditional density. As an application, we consider the case where $\Theta$ is the space of q-smooth functions on $[0,1]^d$ metrized with the $L_r$-distance, $1 \leq r < \infty$. An upper $d$-error bound for Minimum Distance Estimate $\hat{\theta}_n$ is obtained, that depends on Kolmogorov entropy of the space $(\Theta, d)$, and holds also under weak dependence; $d$ the $L_1$-distance. It is risk optimal, when $\Theta$ is the space of $q$-smooth functions on $[0,1]^d$.

The tools needed to obtain the results are presented in the sequel.

# 2 Optimal Estimates in Probability and in Risk

In the previous lectures we obtained the upper convergence rate of the MDE in Probability. The question that arises is whether the estimate and the rate are optimal. Definitions of optimality for a sequence of estimates $\{\hat{\theta}_n\}$ in Probability and in Risk follow.

Let $\{\hat{\theta}_n, \hat{T}_n\}$ be estimates of the parameter $\theta(\in \Theta)$. The estimation loss/cost is measured by a distance measure $d$. The observations $X_1, \ldots, X_n$ are *i.i.d.* with respect to probability measure $P_\theta, \theta \in \Theta$.

For a sequence of estimates $\{\hat{\theta}_n\}$ of $\theta$ it is expected that the Risk

$$E_\theta d(\hat{\theta}_n, \theta) \underset{n\to\infty}{\longrightarrow} 0 \quad \forall \quad \theta \in \Theta,$$

or, for the convergence of the risk to zero to be independent of $\theta$,

$$\sup_{\theta\in\Theta} E_\theta d(\hat{\theta}_n, \theta) \leq C_U \gamma_n; \quad \gamma_n \downarrow 0, \quad C_U = \text{constant}.$$

We would feel more comfortable if there is no other sequence of estimates $\{\hat{T}_n\}$ that converges faster than $\{\hat{\theta}_n\}$ uniformly in $\theta$, i.e.

$$\inf_{\hat{T}_n} \sup_{\theta\in\Theta} E_\theta d(\hat{T}_n, \theta) \geq C_L \gamma_n; \quad \gamma_n \downarrow 0, \quad C_L = \text{constant}$$

These observations motivate the definition of finite sample risk optimality.

**Definition 2.1** *A sequence of estimates* $\{\hat{\theta}_n\}$ *is risk-optimal in estimating* $\theta \in (\Theta, d)$ *with rate of convergence* $a_n(\downarrow 0)$, *if there are constants* $C_L$ *and* $C_U$, $0 < C_L \leq C_U$, *such that for* $n \geq 1$,

$$C_L \cdot a_n \leq \underbrace{\inf_{\hat{T}_n} sup_{\theta\in\Theta} Ed(\hat{T}_n, \theta)}_{\text{Minimax Risk}} \leq sup_{\theta\in\Theta} Ed(\hat{\theta}_n, \theta) \leq C_U \cdot a_n. \tag{3}$$

For the rate of convergence in Probability of $\{\hat{\theta}_n\}$ to $\theta$ consult the definition of achievable $d$-upper rate of convergence given in previous lectures that provides also (5).

**Definition 2.2** $\{\hat{\theta}_n\}$ *is d-optimal in probability uniformly in* $\theta$ *if there is a sequence* $\{a_n\}$ *decreasing to zero such that*

$$\lim_{C\to 0} \liminf_{n\to\infty} \inf_{\hat{T}_n} \sup_{\theta\in\Theta} P[d(\hat{T}_n, \theta) > Ca_n] = 1 \tag{4}$$

*and*

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{\theta\in\Theta} P[d(\hat{\theta}_n, \theta) > Ca_n] = 0. \tag{5}$$

Density estimates satisfying (5) have been already obtained, with the achievable upper rate of convergence $a_n$ depending on the Kolmogorov entropy of the parameter space. The same will hold also for regression estimates and the upper bound in (3). The lower bounds for the convergence rates in (3) and (4) depend on the behavior of the Kullback-Leibler information of the underlying probability measures of the sample,"locally" at each $\theta \in \Theta$.

# 3 Inequalities

Convexity is used repeatedly to obtain inequalities. The material is presented for completeness.

**Definition 3.1** *Let* $f(x)$ *be a real valued function defined on the interval* $I = [a, b]$. $f$ *is* <u>convex</u> *if for every* $x_1, x_2 \in [a, b]$ *and* $0 \leq p \leq 1$,

$$f[px_1 + (1-p)x_2] \leq pf(x_1) + (1-p)f(x_2). \tag{6}$$

• Draw a graph to see what (6) means.

Informally: $f$ is convex when for every segment $[x_1, x_2]$, as $x_p = px_1 + (1-p)x_2$ varies over the line segment $[x_1, x_2]$, the point $(x_p, f(x_p))$ lies below the segment connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$.

• $f$ is strictly convex if inequality (6) is strict for $x_1 \neq x_2$.

• If $f''(x) \geq 0 \ \forall \ x \in [a, b]$ then $f$ is convex for all $x \in [a, b]$.

**Definition 3.2** *$f$ is concave (strictly concave) if $(-f)$ is convex (strictly convex)*

**Examples:** $f_1(x) = x^2, g_1(x) = e^x$ are convex; $f_2(x) = -x^2, g_1(x) = \log x$ are concave.

**Remark 3.1** Let $h(x) = x \log x$. Is $h(x)$ convex or concave? $h(x)$ is used repeatedly to prove inequalities.

**Jensen's inequality (for discrete r.v.'s)**: Let $x_1, \ldots, x_k$ be in the Interval $I$, $0 \leq p_i \leq 1, \sum_{i=1}^k p_i = 1$. If $f$ is convex, then

$$f(\sum_{i=1}^k x_i p_i) \leq \sum_{i=1}^k p_i f(x_i). \tag{7}$$

• Result (7) says:

- if $f$ is convex, $f(EX) \leq Ef(X)$;
- if $g$ is concave, $Eg(X) \leq g(EX)$.

• When you have to prove inequalities using Jensen's inequality the function that should play the role of $f$ and $g$ may not be easy to recognize. An application of Jensen's inequality that will be used later follows.

**Lemma 3.1** *Let $a_i > 0, b_i > 0, i = 1, \ldots, n, a = \sum_{i=1}^n a_i, b = \sum_{i=1}^n b_i$.*

*a) Show that*

$$a \cdot \log \frac{a}{b} \leq \sum_{i=1}^n a_i \cdot \log \frac{a_i}{b_i}. \tag{8}$$

*(Hint: You may use Remark 3.1.)*

*b) Assume in addition that $a_i = c > 0, \bar{b} = \frac{1}{n} \sum_{i=1}^n b_i$, then*

$$nc \cdot \log \frac{nc}{\sum_{i=1}^n b_i} \leq \sum_{i=1}^n c \log \frac{c}{b_i} \longrightarrow c \cdot \log \frac{c}{\bar{b}} \leq \frac{1}{n} \sum_{i=1}^n c \log \frac{c}{b_i}. \tag{9}$$

**Proof of Lemma 3.1 :** *a)* In the right hand side of (8) the $\log(x)$ is a concave function and if Jensen would be used the inequality should have been reversed. If $x\log(x)$ would be the function then the inequality would be of this type BUT we have to alter the presentation for this sum:

$$\sum_{i=1}^{n} a_i \cdot \log \frac{a_i}{b_i} = \sum_{i=1}^{n} b_i \frac{a_i}{b_i} \cdot \log \frac{a_i}{b_i}$$

now we get the function $f(x) = x\log(x)$ but it is multiplied by $b_i$ which is not a probability. Then, we can make it by rewriting

$$\sum_{i=1}^{n} a_i \cdot \log \frac{a_i}{b_i} = \sum_{i=1}^{n} b_i \frac{a_i}{b_i} \cdot \log \frac{a_i}{b_i} = b \sum_{i=1}^{n} \frac{b_i}{b} \cdot \{\frac{a_i}{b_i} \cdot \log \frac{a_i}{b_i}\} \geq bf(\sum_{i=1}^{n} \frac{b_i}{b} \frac{a_i}{b_i}) = bf(\frac{a}{b}) = b\frac{a}{b}\log(\frac{a}{b})$$

*b)* Follows from

**Proof of Jensen's inequality for discrete r.v.:** For $k = 1, 2$, (7) holds by convexity. Assume (7) holds for $k = m$, i.e.

$$f(\sum_{i=1}^{m} p_i x_i) \leq \sum_{i=1}^{m} p_i f(x_i).$$

To show (7) holds for $k = m + 1$ we work for the right side of the inequality:

$$\sum_{i=1}^{m+1} p_i f(x_i) = \sum_{i=1}^{m} p_i f(x_i) + p_{m+1} f(x_{m+1}) = (1 - p_{m+1}) \sum_{i=1}^{m} \frac{p_i}{1 - p_{m+1}} f(x_i) + p_{m+1} f(x_{m+1})$$

$$\geq (1 - p_{m+1}) f(\sum_{i=1}^{m} \frac{p_i x_i}{1 - p_{m+1}}) + p_{m+1} f(x_{m+1}) \geq f(\sum_{i=1}^{m+1} p_i x_i).$$

- Equality in (7) holds iff f is linear or $x_1 = \ldots = x_k$.

**Corollary 3.1** *Let* $p_i > 0 : \sum_{i=1}^{N} p_i = 1$, *then*

$$\sum_{i=1}^{N} p_i \log \frac{1}{p_i} \leq \log \sum_{i=1}^{N} p_i \frac{1}{p_i} = \log N. \tag{10}$$

**Observe that the upper bound is achieved when** $p_i = \frac{1}{N}, i = 1, \ldots, N$, **to be used to prove the last inequality in Fano's Lemma.**

**Proposition 3.1** *If* $f$ *is convex on* $[a, b]$ *then it takes values larger than the tangent line at any* $x_0 \in (a, b)$, *i.e.*

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0). \tag{11}$$

**Proof of Proposition 3.1:** Let $x_1 < x_2 < x_3, \longrightarrow x_2 = px_3 + (1-p)x_1 \longrightarrow p = \frac{x_2-x_1}{x_3-x_1}, 1-p = \frac{x_3-x_2}{x_3-x_1}$.

From convexity of $f$, $\frac{f(x_2)-f(x_1)}{x_2-x_1} \leq \frac{f(x_3)-f(x_2)}{x_3-x_2}$.

Indeed,

$$f(x_2) = f(px_3+(1-p)x_1) \leq pf(x_3)+(1-p)f(x_1) \longrightarrow 0 \leq p[f(x_3)-f(x_2)]+(1-p)[f(x_1)-f(x_2)]$$

$$\longrightarrow [f(x_2)-f(x_1)](x_3-x_2) \leq [f(x_3)-f(x_2)](x_2-x_1) \longrightarrow \frac{f(x_2)-f(x_1)}{x_2-x_1} \leq \frac{f(x_3)-f(x_2)}{x_3-x_2}.$$

Let now

$$x_2 = x_1 + h \longrightarrow \frac{f(x_1+h)-f(x_1)}{h} \leq \frac{f(x_3)-f(x_1+h)}{x_3-x_1-h}.$$

Taking limits in both sides of the inequality as $h \to 0$

$$f'(x_1) \leq \frac{f(x_3)-f(x_1)}{x_3-x_1} \longrightarrow f(x_3) \geq f(x_1) + f'(x_1)(x_3-x_1)$$

proving (11).

Let now

$$x_2 = x_3 - h \longrightarrow \frac{f(x_3-h)-f(x_1)}{x_3-h-x_1} \leq \frac{f(x_3)-f(x_3-h)}{h}.$$

Taking limits in both sides of the inequality as $h \to 0$

$$\frac{f(x_3)-f(x_1)}{x_3-x_1} \leq f'(x_3) \longrightarrow f(x_1) \geq f(x_3) + f'(x_3)(x_1-x_3)$$

proving (11).

- Use of Proposition 3.1 to prove the general case of Jensen's inequality.

**Proposition 3.2** *Let $X$ be a r.v. with domain the real line and with expected value $EX$. Let $f$ be a convex function with domain the range of the values of $X$. Then,*

$$f(EX) \leq Ef(X). \tag{12}$$

**Remark 3.2** *To use (12) when there is an integral $\int_A f(x)dx$ rewrite the integral as $\int I_A(x)f(x)dx$ and use the inequality for the function $I_A(x)f(x)$; $I_A(x) = 1$ if $x \in A$ and 0 otherwise.*

**Proof of Proposition 3.2:** Let consider the tangent line, $L(x)$, to $f$ at $(EX, f(EX))$,

$$L(x) = f(EX) + f'(EX)(x - EX).$$

Then, from (11) it follows that

$$f(X) \geq L(X) \longrightarrow Ef(X) \geq EL(X) = f(EX).$$

$x_1, \ldots, x_m$ non-negative reals

$$\sum_{i=1}^{m} x_i \log x_i \geq t \log t + (m - t) \log \frac{m - t}{m - 1} \geq m \log \frac{m}{2} - (m - t) \log m$$

$t = \max\{x_i\}, m = \sum_{i=1}^{m} x_i$

Distances and deviations between Probability measures/densities

Let $P, Q$ measures on a space $\mathcal{X}$ with a $\sigma$-field $\mathcal{A}$. Assume the measures have densities $p$ and $q$ respectively, with respect to dominating measure$\mu$ : $\frac{dP}{d\mu} = p$, $\frac{dQ}{d\mu} = q$. You can think of $\mu$ as Lebesgue measure, i.e. $\mu(dx) = dx$.

- $L_1$-distance (or Total Variation distance) between $P, Q$ :

$$||P - Q||_1 = 2 \sup_{A \in \mathcal{A}} |P(A) - Q(A)| \tag{13}$$

- It holds:

$$||P - Q||_1 = 2[P(x : p(x) > q(x)] - Q[x : p(x) > q(x)] = \int_{\mathcal{X}} |p(x) - q(x)| \mu(dx). \tag{14}$$

- Kullback-Leibler non-distance (WHY?) between $P, Q$ :

$$d_{KL}(P, Q) = d_{KL}(p, q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mu(dx)$$

- $d_{KL}(P, Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mu(dx) = - \int_{\mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \mu(dx) = E_P[$ **- log**$\frac{q(X)}{p(X)}]$
$\geq$ **-log**$E_P \frac{q(X)}{p(X)} = -\log(1) = 0$

- $L_r$**-distance:** $||p - q||_r = \left[\int_{\mathcal{X}} |p(x) - q(x)|^r \mu(dx)\right]^{1/r}, 1 \leq r < \infty$.

- $L_\infty$**-distance:** $||p - q||_\infty = \sup_{x \in \mathcal{X}} |p(x - q(x)|$.

# 4 Shannon's Entropy, Information, Fano's Lemma

**Entropy/Uncertainty of discrete r.v.**

Shannon (1948) wanted a measure $H$ of uncertainty/surprise of events with certain properties. Before him a definition was given by Hartley for equally likely events. Some required properties for $H$ :

$$H(certainty) = H(no\ uncertainty) = H(no\ surprise) = 0;$$

$H(largest\ uncertainty/surprise)$ will give the highest $H$-value;

additional information reduces the uncertainty.

The events are outcomes of a discreet random variable. Shannon's related work started in Bell laboratories using a binary system, i.e. 0's and 1's to describe the results of coin tossing, i.e. $H$(eads) and $T$(ails). Instead of calling each digit a "binary digit" John Tukey suggested the word "bit". Shannon suggested for $H$ a measure used in Statistical Mechanics which for coin tossing, with $p =$ Probability of Head and log denoting $\log_2$ is

$$-[p \log p + (1-p) \log(1-p)].$$

Observe that when $p = 1$, i.e. complete certainty of the result

$$-[1 \cdot \log 1 + 0 \log 0] = 0,$$

if $0 \cdot (-\infty) = 0$. The most uncertain situation for the coin tossing is when $p = 1/2$ and

$$-[\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}] = \log_2 2 = 1,$$

which either using Jensen's inequality or simply finding at the extremum of the function $p \log_2 p + (1-p) \log_2(1-p)$ is indeed attained at $p = \frac{1}{2}$.

**Definition 4.1** *Let $X$ be a discrete r.v., $p(x) = P(X = x), x \in \mathcal{X}$. Then its entropy*

$$H(X) = -E \log p(X) = E \frac{1}{\log p(X)} = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x). \tag{15}$$

*It follows that for discrete r.v. $X, Y$ with joint probability $p_{X,Y}$ their entropy*

$$H(X,Y) = -E \log p_{X,Y}(X,Y) = E \log \frac{1}{p_{X,Y}(X,Y)} = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{X,Y}(x,y) \log p_{X,Y}(x,y). \tag{16}$$

The notation $H(X|Y)$ is used for what is called "average posterior entropy". Do not confuse $H(X|Y)$ with conditional probability statements: $H(X|Y)$ is a number.

**Definition 4.2** *For discrete r.v. $X, Y$,*

$$H(X|Y) = -E \log p(X|Y) = E \frac{1}{\log p(X|Y)} = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{X,Y}(x,y) \log p_{X|Y}(x|y). \tag{17}$$

• **However,** the entropy $H(X)$ is an expected value thus the identity for expected value $E(V) = E[E(V|U)]$ holds, i.e.

$$H(X) = -E\log p(X) = -E[E(\log p(X)|U)] = -\sum_{u\in\mathcal{U}} E(\log p(X)|U = u)\cdot P(U = u). \quad (18)$$

**Properties of $H$**

**a**) If the cardinality of $X$-values is $N$ then, from (10), $H(X) \leq \log N$. The upper bound is achieved when values of $X$ are equally likely.

**b**) $H(X,Y) = H(X) + H(Y|X)$

**Proof:** $H(X,Y) = -E\log p(X,Y) = -E\log p(X)p(Y|X) = -E\log p(X) - E\log p(Y|X)$.

**c**) $H(X,Y) \leq H(X) + H(Y)$

**Proof:** Use $r_{i,j} = P(X = x_i, Y = y_j), p_i = P(X = x_i), q_j = P(Y = y_j)$.
Observe that $\sum_j r_{i,j} = p_i$ and similarly for $q_j$.
$H(X,Y) - H(X) - H(Y) = \sum_{i,j} r_{i,j}\log\frac{1}{r_{i,j}} + \sum_i p_i\log p_i + \sum_j q_j\log q_j = \sum_{i,j} r_{i,j}\log\frac{1}{r_{i,j}}$
$+\sum_{i,j} r_{i,j}\log p_i + \sum_{j,i} r_{i,j}\log q_j = \sum_{i,j} r_{i,j}\log\frac{p_iq_j}{r_{i,j}} \leq (Why?)\log\sum_{i,j} p_iq_j = 0.$

**d**) $H(X,Y) = H(X) + H(Y) \iff X, Y$ are independent.

**Proof:** log is not linear thus equality holds in **c)** holds iff $\frac{r_{i,j}}{p_iq_j} = k,\ \forall\ i,j$, or $r_{i,j} = kp_iq_j$ and summing for all $j$ in both sides we get $p_i = kp_i$ or $k = 1$ or $X, Y$ are independent.

**e**) From **c**), for $U_1,\ldots,U_n$ discrete r.vs, $H(U_1,\ldots,U_n) \leq H(U_1) + \ldots + H(U_n)$.

**f**) $H(X|Y) \leq H(X)$ (Higher surprise if nothing is known.)

**Proof:** From **b)** and **c)**, respectively,

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y),$$

$$H(X,Y) \leq H(X) + H(Y),$$

therefore, $H(X|Y) \leq H(X)$.

**g**) For any function $g,\ H(g(Y)|Y) = 0$.

Intuitively clear: when $Y$ is known there are no surprises for $g(Y)$. In the calculations $P[g(Y) = g(y)|Y = y]$ is needed. What is its value?

**h)** $H(Y|X) \leq H(Y|g(X))$ Use $G = g(X)$ with $G$ taking values $g$ (abuse of notation). Intuitively: $g(X)$ shrinks the $X$-values, so there will be more surprise when you know $g(X)$ instead of $X$.

**Proof:** $H(Y|X) = -\sum_y \sum_x p_{Y,X}(y,x) \log \frac{p_{Y,X}(y,x)}{p_X(x)}$

$$H(Y|G) = -\sum_y \sum_g p_{Y,G}(y,g) \log \frac{p_{Y,G}(y,g)}{p_G(g)}$$

We will use repeatedly that for fixed $y$ :

$$p_{Y,G}(y,g) = P[Y=y, G=g] = P[\cup_{\{x:g(x)=g\}}\{Y=y, X=x\}] = \sum_{x:g(x)=g} p_{Y,X}(y,x)$$

We work on the term $p_{Y,G}(y,g) \log \frac{p_{Y,G}(y,g)}{p_G(g)}$ and use that in (8), with $a = \sum a_i,\ b = \sum b_i, a_i > 0, b_i > 0$,

$$a \cdot \log \frac{a}{b} \leq \sum_{i=1}^{n} a_i \cdot \log \frac{a_i}{b_i},$$

$$p_{Y,G}(y,g) \log \frac{p_{Y,G}(y,g)}{p_G(g)} = [\sum_{x:g(x)=g} p_{Y,X}(y,x)] \log \frac{\sum_{x:g(x)=g} p_{Y,X}(y,x)}{\sum_{x:g(x)=g} p_X(x)} \leq \sum_{x:g(x)=g} p_{Y,X}(y,x) \log \frac{p_{Y,X}(y,x)}{p_X(x)}$$

Thus,

$$H(Y|G=g(X)) = -\sum_y \sum_g p_{Y,G}(y,g) \log \frac{p_{Y,G}(y,g)}{p_G(g)}$$

$$\geq -\sum_y \sum_g \sum_{x:g(x)=g} p_{Y,X}(y,x) \log \frac{p_{Y,X}(y,x)}{p_X(x)} = H(Y|X)$$

**Proposition 4.1** *(Fano's Inequality) Let $X, Y$ be r.v. with values $1, \ldots, N$, $W = I_{\{X \neq Y\}}$ is the indicator of $\{X \neq Y\}$, with value 1 when $X \neq Y$ and value 0 when $X = Y$. Then,*

$$H(X|Y) \leq \log 2 + P(X \neq Y) \log(N-1). \tag{19}$$

**Proof:** From **b)** of the $H$-properties, $H(U,V) = H(U) + H(V|U)$, and therefore

$$H(W,X|Y) = H(X|Y) + H(W|X,Y) = H(X|Y) \tag{20}$$

with the last equality due to **g)** since $W$ is function of $X,\ Y$.

Reversing the roles of $W,\ X$ in the middle equality in (20),

$$H(W,X|Y) = H(W|Y) + H(X|W,Y)$$

observing that $W$ gets only 2 values and using from **a)** the bound on $H(W)$ and (18),

$$\leq \log 2 + P(W=1)H(X|W=1,Y) + P(W=0)H(X|W=0,Y)$$

$$= \log 2 + P(W=1)\log(N-1) + 0 = \log 2 + P(X \neq Y)\log(N-1)$$

where the penultimate equality holds since
*i)* when $W=1$ then $X \neq Y \rightarrow X|\{W=1,Y\}$ takes $N-1$ values,
*ii)* when $W = I_{\{X \neq Y\}} = 0$ it means $X = Y$ and therefore $X|\{W=0,Y\}$ takes the $Y$-value and there are no surprises *i.e.* $H(X|W=0,Y)=0$.

**Information $I(X,Y)$ in random variables $X,Y$ via $d_{KL}$**

$I(X,Y) = d_{KL}(P_{X,Y}, P_X x P_Y)$

Thus, $I(X,Y)$ is a measure of the difference between $p_{X,Y}(x,y)$ and $p_X(x)\cdot p_Y(y)$ on the average with respect to the joint probability. We will see below that this difference is equal also with the difference in uncertainties/surprises between $X$ and $X|Y$, and if $X|Y$ takes $N$ values then

$$I(X,Y) = d_{KL}(P_{X,Y}, P_X x P_Y) = H(X) - H(X|Y) \leq \log N - H(X|Y). \tag{21}$$

**Proof:** $I(X,Y) = d_{KL}(P_{X,Y}, P_X x P_Y) = \sum_{(x,y)\in \mathcal{X}x\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$
$= \sum_{(x,y)\in \mathcal{X}x\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X|Y}(x|y)}{p_X(x)} = \sum_{x\in\mathcal{X}}[\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)]\log\frac{1}{p_X(x)}$
$-[-\sum_{(x,y)\in \mathcal{X}x\mathcal{Y}} p_{X,Y}(x,y)\log p_{X|Y}(x|y)] = H(X) - H(X|Y)$.
The upper bound follows from (10) and it is achieved when $X$ takes each of the $N$-values with probability $1/N$.

**Fano's Lemma**

Recall Fano's inequality (19),

$$H(X|Y) \leq \log 2 + P(X \neq Y)\log(N-1). \tag{22}$$

which implies using also (21) with $X$'s values equally likely,

$$P(X \neq Y) \geq \frac{H(X|Y) - \log 2}{\log(N-1)} = \frac{H(X) - I(X,Y) - \log(2)}{\log(N-1)} = \frac{\log(N) - I(X,Y) - \log(2)}{\log(N-1)} \tag{23}$$

or,

$$P(X \neq Y) \geq 1 - \frac{I(X,Y) + \log 2}{\log(N-1)}. \tag{24}$$

**Use of (23) in estimation:** $X, Y$ be discrete random variables, $Y$ uniform taking $N$ values,

$$P(X = x|Y = i) = p_i(x), \qquad P(X = x, Y = i) = p_i(x)/N.$$

$Y$ determines the $N$ potential models followed by $X$ given $Y$.

$$I(X,Y) = \sum_{i=1}^{N} \sum_{x} p(X = x, Y = i) \log \frac{P(X = x, Y = i)}{P(X = x)P(Y = i)}$$

$$= \sum_{i=1}^{N} \sum_{x} \frac{p_i(x)}{N} \log \frac{p_i(x)/N}{\sum_{j=1}^{N} p_j(x)/N^2} = \frac{1}{N} \sum_{i=1}^{N} \sum_{x} p_i(x) \log \frac{p_i(x)}{\sum_{j=1}^{N} p_j(x)/N}$$

Work on the term in the sum, looking at $p_i$ as constant, $c$, and use (9):

$$p_i(x) \log \frac{p_i(x)}{\sum_{j=1}^{N} p_j(x)/N} = \frac{1}{N} \cdot N p_i(x) \log \frac{N p_i(x)}{\sum_{j=1}^{N} p_j(x)}$$

$$\leq \frac{1}{N} \sum_{j=1}^{n} p_i(x) \log \frac{p_i(x)}{p_j(x)}$$

thus an upper bound on the information is obtained for these $X, Y$

$$I(X,Y) \leq \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{KL}(p_i, p_j) \tag{25}$$

and from (24)

$$P(X \neq Y) \geq 1 - \frac{N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{KL}(p_i, p_j) + \log 2}{\log(N-1)}. \tag{26}$$

Observe that:

$$P(X \neq Y) = \sum_{i=1}^{N} P(X \neq i|Y = i)P(Y = i) = \frac{1}{N} \sum_{i=1}^{N} P(X \neq i|Y = i). \tag{27}$$

In estimation problems, let $(\mathcal{X}, \mathcal{A})$ be a space with a $\sigma$-field, $\mathbf{X}$ is a random vector on $\mathcal{X}$ obtained from one of the probability measures $P_1, \ldots, P_N$. Let $Y$ be a uniform r.v. taking values $1, \ldots, N$ and $\delta(\mathbf{X})$ be an estimate of the index of the probability of $\mathbf{X}$. We are interested in

$$P[\mathbf{X} : \delta(\mathbf{X}) \neq i||Y = i] = P_i[\mathbf{X} : \delta(\mathbf{X}) \neq i], i = 1, \ldots, N,$$

in particular in their average which will be used to find a lower bound on the $L_r$-error, $r \geq 1$, via (26) and (27)

$$\frac{1}{N}\sum_{i=1}^{N} P_i[\mathbf{X} : \delta(\mathbf{X}) \neq i] = P[\delta \neq Y] \geq 1 - \frac{N^{-2}\sum_{i,j} d_{KL}(P_i, P_j) + \log 2}{\log(N-1)}; \qquad (28)$$

$P_i$ in (28) denotes the joint probability of $\mathbf{X}$, $i = 1, \ldots, n$.

# 5 Lower Bounds on the Error in Nonparametric Regression-Type Problems

In the classical nonparametric regression problem, we consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_1, \ldots, X_n$ are $R^d$ valued measurements that might be random or nonrandom, $Y_1, \ldots, Y_n$ are the corresponding responses such that $E(Y_i|X_i = x_i) = \theta(x_i)$ with $\theta$ in an infinite dimensional space $\Theta$. Conditionally on $X_1 = x_1, \ldots, X_n = x_n$, the $Y$-responses are independent with distributions of the same form $f(y|x, \theta(x)dy := P_{\theta(x)}(dy)$, but with parameters depending on the measurements $x_i, i = 1, \ldots, n$.

Under this setup, Stone (1980, 1982) and Ibragimov and Khas'minskii (1980) have constructed optimal estimators $\hat{\theta}_n$ of $\theta$ in $L_r, 1 \leq r$, when $\Theta$ consists of $q$-smooth functions on $[0,1]^d$. Ibragimov and Khas'minskii proved that their estimators are almost minimax modulo a constant, that is, there are constants $C_L, C_U$ such that $\sup\{E_\theta||\hat{\theta}_n - \theta||_r; \theta \in \Theta\} \leq C_U \cdot n^{-\gamma}$ and $\inf\{\sup\{E_\theta||\hat{T}_n - \theta||_r; \theta \in \Theta\}; \hat{T}_n\} \geq C_L \cdot n^{-\gamma}, \gamma > 0, n \in N$. Stone has considered other definitions of optimality using bounds for the loss $||\hat{\theta}_n - \theta||_r$ as described at the end of the paper.

We will relax the condition that $\theta(x)$ is a conditional mean as in the classical regression problem. We will only assume that $\theta(x)$ is a parameter of the conditional density and we will call the problem of estimating $\theta$ a regression type problem. When $\theta$ is an element of a metric space $(\Theta, d)$ we will provide for the regression type problem a lower bound on the d-minimax risk.This bound can be used as a tool to provide lower bounds for different choices of $(\Theta, d)$. We will apply the result and evaluate the lower bound in the case where $\Theta$ is a family of smooth functions and $d$ is the $L_r$- distance, $1 \leq r < \infty$. In a remark

at the end of the paper, we provide with the same technique, lower bounds for the $d$-loss in probability. A minimum distance estimate for the regression type problem proposed in Yatracos (1989) shows achievability of the lower bound. The error of this estimate depends on the entropy of Kolmogorov of $(\Theta, d)$ as in the density estimation problem (Yatracos, 1985).

The method for computing the lower bound comes from Le Cam's idea in hypothesis testing (Le Cam, 1986, or Kraft,1955) that you cannot test (and so estimate) $\theta_0$ versus $\Theta - \{\theta_0\}$ if $\theta_0$ is in the convex hull of $\Theta - \{\theta_0\}$. So it will be difficult to test $\theta_0$ versus $\Theta_n \subseteq \Theta - \{\theta_0\}$ when $\Theta_n$ consists of functions close to $\theta_0$, the difficulty being reflected in the lower bound of the minimax or Bayes risk. This idea has already been used by Bretagnolle and Huber (1979) to obtain lower bounds for the risk in the nonparametric density estimation problem. A similar approach, using Fano's lemma, has been considered to obtain lower bounds for minimax risks by Khas'minskii (1978) and Birge (1983) in density estimation and by Ibragimov and Khas'minskii (1981) in classical regression with equidistant design. An observation that a regression problem is almost a density estimation problem leads to the use of Fano's lemma and a lower bound for an arbitrary metric space $(\Theta, d)$. An elegant result of Birgé helps to obtain the best lower bound when $\Theta$ is the space of $(q, L)$ smooth functions on $[0,1]^d$ metrized with the $L_r$ distance, $1 \leq r < \infty$, (i.e., $\Theta$ consists of $p$ times differentiable functions in $[0,1]^d$, uniformly bounded in sup-norm with the $p-th$ derivative satisfying a Lipschitz condition with parameters $(L, a), q = p + a, 0 \leq p, 0 < a < 1$). Note that Fano's lemma involves the Kullback information $K(P_{\theta_1(x)}, P_{\theta_2(x)})$, so we will have to evaluate it or find an upper bound for it. It is easy to see that for the case considered by Stone, $K(P_{\theta_1(x)}, P_{\theta_2(x)}) \leq C[\theta_1(x) - \theta_2(x)]^2$. It is this condition that makes the estimators of Ibragimov and Khas'minskii and Stone asymptotically optimal and not the nature of $\theta(x)$ in the conditional density. It is the behavior of the Kullback information $K(P_{\theta_1(x)}, P_{\theta_2(x)})$ locally that will determine the lower bound on the risk and the lower bound of the loss in probability.

For sample size $n$, we will compute a lower bound on the $\sup\{E_\theta d(\hat{T}_n, \theta); \theta \in \Theta\}$ by considering a bound on $\sup\{E_\theta d(\hat{T}_n, \theta); \theta \in \Theta_n\}$ with $\Theta_n$ an appropriate subset of $\Theta$ according to Le Cam's idea. It turns out that when $\Theta$ is the set of $q$-smooth functions

on $[0,1]^d$, one can use a set $\Theta_n$ similar to the one used by Kolmogorov and Tihomirov (1959) to compute a lower bound for the entropy of smooth functions on $[0,1]^d$. We should also mention, at this point, the work by Boyd and Steele (1978) and Assouad (1983). The former proved that in the nonparametric density estimation problem, considering all densities with squared error loss, the minimax risk cannot be better than $O(n^{-1})$. The latter provided a lower bound on risks for any loss and related the $O(n^{-1/2})$ minimax risk with dimensionality properties of the space of probability measures under consideration.

Khas'minskii (1978) provides lower bounds on the risks of nonparametric estimates of densities in the uniform metric. Devroye (1986) computes minimax bounds on the $L_1$ loss for the class of kernel estimates. For a detailed study on lower bounds on minimax risks, the reader could consult Devroye and Gyorfi (1985) and Devroye (1987).

**Notation. Definitions. The results.** Let $(\mathcal{X},\mathcal{A}),(\mathcal{Y}_1,\mathcal{B}_1),\ldots,(\mathcal{Y}_n,\mathcal{B}_n)$ be spaces with their $\sigma$-fields. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a sample with $X_i$ taking values in $\mathcal{X}, i=1,\ldots,n, Y_i$ taking values in $\mathcal{Y}_i, i=1,\ldots,n$.

**Definition 5.1** *For any two functions $f,g$ on $(\mathcal{X},\mathcal{A})$, $L_r(\lambda)$ integrable, $1\le r<\infty$, their $L_r$ distance is*

$$||f-g||_r=\left(\int_{\mathcal{X}}|f(x)-g(x)|^r\lambda(dx)\right)^{1/r}. \tag{29}$$

**Definition 5.2** *For any two probability measures $P,Q$ on $(\mathcal{Y},\mathcal{B})$, their Kullback information $K(P,Q)=E_P\log(\frac{dP}{dQ})$ if $P$ is absolutely continuous with respect to $Q$; otherwise, $K(P,Q)=\infty$.*

**Note:** In the case of product measures $K(P_1xP_2x\ldots xP_n,Q_1xQ_2x\ldots Q_n)=\sum_{i=1}^n K(P_i,Q_i)$.

**FANO'S LEMMA:** Let $(\mathcal{Y},\mathcal{B})$ be a space with a $\sigma$-field, $P_1,\ldots,P_m$ probability measures on $\mathcal{B}$ and $\delta$ an estimator of the measures defined on $\mathcal{Y}$. Then

$$\frac{1}{m}\sum_{i=1}^m P_i[\delta(y)\neq P_i]\ge 1-\frac{m^{-2}\sum_i\sum_j K(P_i,P_j)+\log 2}{\log(m-1)}. \tag{30}$$

**Proposition 5.1** *Under the regression setup used for the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$, for $\Theta_n$ a subset of $\Theta$ with finite cardinality, $d$ a distance on $\Theta$ and $\hat{T}_n$ an estimator of the*

*regression type function* $\theta$,

$$\sup\{E_\theta d(\hat{T}_n, \theta) : \theta \in \Theta\}$$

$$\geq \frac{1}{2}\inf\{d(\theta_1,\theta_2) : \theta_1 \neq \theta_2, (\theta_1,\theta_2) \in \Theta_n^2\} E\left[1 - \frac{(card(\Theta_n))^{-2}\sum_{(\theta_1,\theta_2)\in\Theta_n^2}\sum_{i=1}^n K(P_{\theta_1(X_i)}, P_{\theta_2(X_i)}) + \log 2}{\log(card(\Theta_n - 1))}\right],$$

*the last expectation taken with respect to the probability measure of* $(X_1, \ldots, X_n)$.

**Proof:** Define $\hat{T}_n^*$ taking values in $\Theta_n$ such that $d(\hat{T}_n, \hat{T}_n^*) = \inf\{d(\hat{T}_n, \theta); \theta \in \Theta_n\}$. Then, we have for $\theta \in \Theta_n$,

$$d(\hat{T}_n^*, \theta) \leq d(\hat{T}_n^*, \hat{T}_n) + d(\hat{T}_n, \theta) \leq 2 \cdot d(\hat{T}_n, \theta)$$

and

$$\sup\{E_\theta d(\hat{T}_n, \theta) : \theta \in \Theta\} \geq \sup\{E_\theta d(\hat{T}_n, \theta) : \theta \in \Theta_n\} \geq \frac{1}{card(\Theta_n)}\sum_{\theta\in\Theta_n} E_\theta d(\hat{T}_n, \theta)$$

$$= E\left[\frac{1}{card(\Theta_n)}\sum_{\theta\in\Theta_n} E_\theta(d(\hat{T}_n, \theta)|X_1 = x_1, \ldots, X_n = x_n)\right]$$

$$\geq \frac{1}{2}\inf\{d(\theta_1,\theta_2); \theta_1 \neq \theta_2, (\theta_1,\theta_2) \in \Theta_n^2\} \cdot \frac{1}{card(\Theta_n)}\sum_{\theta\in\Theta_n} P_\theta[\hat{T}_n^* \neq \theta | X_1 = x_1, \ldots, X_n = x_n]$$

$$\geq \frac{1}{2}\inf\{d(\theta_1,\theta_2); \theta_1 \neq \theta_2, (\theta_1,\theta_2) \in \Theta_n^2\}\left[1 - \frac{(card(\Theta_n))^{-2}\sum_{(\theta_1,\theta_2)\in\Theta_n^2}\sum_{i=1}^n K(P_{\theta_1(x_i)}, P_{\theta_2(x_i)} + \log 2}{\log(card(\Theta_n) - 1)}\right],$$

by applying Fano's Lemma (30) to the probability measures $P_{\theta(x_1)}x \ldots x P_{\theta(x_n)}, \theta \in \Theta_n$, on the product spaces $\mathcal{Y}_1 x \ldots x \mathcal{Y}_n$. □

**Corollary 5.1** *Under the assumptions of Proposition 5.1, if* $K(P_{\theta_1(x_i)}, P_{\theta_2(x_i)}) \leq c_n$ *for all* $\theta_1, \theta_2 \in \Theta_n$ *and* $\inf\{d(\theta_1,\theta_2); \theta_1 \neq \theta_2, (\theta_1,\theta_2) \in \Theta_n^2\} > a_n$, *then*

$$\sup\{E_\theta d(\hat{T}_n, \theta) : \theta \in \Theta\} \geq \frac{1}{2}a_n\left[1 - \frac{nc_n + \log 2}{\log(card(\Theta_n) - 1)}\right]. \tag{31}$$

**Proposition 5.2** *Assume for the conditional density* $f(y|x,t)$, *with all derivatives taken with respect to the parameter* $t$, *that:*

*(i)* $\int_{\mathcal{Y}} f'(y|x,t)\mu(dy) = 0$.

*(ii) If* $l(y|x,t) = \log f(y|x,t)$, *there are positive constants* $\epsilon_0$ *and* $K_1$ *and a function*

$M(y|x,y)$ *such that*

**a)** $|l''(y|x,t+\epsilon)| \le M(y|x,t)$ *for* $|\epsilon| \le \epsilon_0$, *and*

**b)** $\int_{\mathcal{Y}} M(y|x,t)\mu(dy) \le K_1$.

*Then,* $K(P_s,P_t) \le C(t-s)^2$.

**Proof:** When $|t-s|$ is small, making a Taylor expansion, we have

$$K(P_s,P_t) = \int_{\mathcal{Y}} f(y|x,t)\log\frac{f(y|x,t)}{f(y|x,s)}\mu(dy)$$

$$= -\int_{\mathcal{Y}} f(y|x,t)\left[(s-t)\frac{f'(y|x,t)}{f(y|x,t)} + \frac{(t-s)^2}{2}l''(y|x,c)\right]\mu(dy) \le \frac{K_1}{2}(t-s)^2,$$

where $c$ is in the open interval determined by $t$ and $s$. $\square$.

**An application**

Let $\Theta$ be the space of $(q,L)$-smooth functions on $[0,1]$. We introduce a family $\Theta_n$ but we will use a subset $\Theta_n^*$ of it to apply Proposition 5.1. Let

$$\phi_{i,n}(x) = ab_n^q\left[1-\left(\frac{x-.5(2i-1)b_n}{.5b_n}\right)^2\right]^q \cdot I_{\{(i-1)b_n \le x \le ib_n\}},\ i=1,\ldots,b_n^{-1},$$

where $a>0$ can be chosen appropriately to make the constant of the Lipschitz condition less than or equal to $L$. The set $\Theta_n$ will consist of functions $\Theta_n$ with form

$$\sum_{i=1}^{b_n^{-1}} \gamma_i\phi_{i,n}(x),$$

with $\gamma_i = 1$ or $0,\ i=1,\ldots,b_n^{-1}$.

Note that the $L_r$ distance between functions of $\Theta_n$ will be greater than or equal to

$$ab_n^q\left[\int_0^{b_n}\left[1-\left(\frac{x-.5b_n}{.5b_n}\right)^2\right]^{qr}\right]^{1/r} = \frac{a}{2^{1/r}}b_n^{q+(1/r)}\left[\int_{-1}^1(1-y^2)^{qr}dy\right]^{1/r} = C_{q,r,a}b_n^{q+\frac{1}{r}}.$$

It is also easy to see that:

$$I_r = \int_{-1}^1(1-y^2)^r dy = \frac{2r}{2r+1}I_{r-1}, r\ge 1,\ I_0 = 2,$$

$$|\theta_1(x)-\theta_2(x)| \le ab_n^q,\ \forall\ \theta_1,\theta_2 \in \Theta_n,\ \forall\ x\in[0,1].$$

When $\mathcal{X} = [0,1]^d$, consider functions $\phi_{j_1,\ldots,j_d,n}$ with form

$$\phi_{j_1,\ldots,j_d,n}(x_1,\ldots,x_d) = ab_n^q \Pi_{i=1}^d \left[1 - \left(\frac{x_i - .5(2j_i-1)b_n}{.5b_n}\right)^2\right]^q \cdot \Pi_{i=1}^d I_{\{(j_i-1)b_n \le x_i \le j_i b_n\}},$$

$\jmath_i = 1,\ldots,b_n^{-1},\ i = 1,\ldots,d.$

Note that there are $b_n^{-d}$ such $I$-rectangles in $[0,1]^d$ defining the $\phi$-functions, so enumerate them as $I_{1,n},\ldots,I_{b_n^{-d},n}$ and the corresponding $\phi$-functions as $\phi_{I_1,n},\ldots,\phi_{I_{b_n^{-d},n}}$. The functions $\theta(x_1,\ldots,x_d)$ in $\Theta_n$ will have form

$$\sum_{i=1}^{b_n^{-d}} \gamma_i \phi_{I_i,n}(x_1,\ldots,x_d), \gamma_i = 0 \text{ or } 1.$$

The lower bound on the $L_r$-distance between functions of $\Theta_n$ will be greated than or equal to $C_{q,r,a,d} b_n^{q+(d/r)}$.

**Definition 5.3** *Let $d$ be a distance-measure on a subset $\mathcal{P}$ of the $L_1(\lambda)$ functions on $(\mathcal{X},\mathcal{A})$, $\lambda$ a probability measure on $\mathcal{A}$, $\Phi$ a function, $\Phi : R^+ \to R^+$. The function $\Phi(d(f,g))$ is called superadditive if for every finite partition $\{A_i; 1 \le i \le l\}$ of $\mathcal{X}$, we have for $f, g$ in $\mathcal{P}$,*

$$\Phi(d(f,g)) = \sum_{i=1}^{l}\sum_{i=1}^{l} \Phi[d(fI_{A_i}, gI_{A_i})]; \tag{32}$$

*$I_A$ denotes the indicator function of $A$.*

Property (32) has been introduced by Bretagnolle and Huber (1979) and is satisfied by $d = ||f-g||_r^r$ on $L_r(\lambda), r \ge 1$.

**Theorem 5.1** *(Birgé, 1983, Proposition 3.8) Let $\{A_i; 1 \le i \le l\}$ be a partition of $\mathcal{X}$ and $f, g_i$ and $g_i'$ be elements of $L_1(\lambda)$ with support on $A_i, i = 1,\ldots,n$. Let $\Theta = \{f + \sum_{i=1}^l \lambda_i; \lambda_i = g_i\}$ and assume that for all $i$, $d(f+g_i, f+g_i')) \ge a$ and that $d^r$ is superadditive for some $r \ge 1$. Then there is a subset $\Theta^*$ of $\Theta$ such that $d(f^*, g^*) \ge a(0.125l)^{1/r}$ for $f^* \ne g^*$ elements of $\Theta^*$ and $\log(card(\Theta^*) - 1) > 0.316l$ for any $l \ge 8$.*

**Corollary 5.2** *If $\Theta$ is the set of $q$-smooth functions on $[0,1]^d$ for conditional densities such that $K(P_{\theta_1(x)}, P_{\theta_2(x)}) \le C|\theta_1(x) - \theta_2(x)|^M$ for some $M > 0$, the $L_r$ minimax risk is greater than or equal to $C^* n^{-q/(Mq+d)}$.*

**Proof:** Consider the preceding sets of functions $\Theta_n$ of the form $\sum_{i=1}^{b_n^{-d}} \gamma_i \phi_{I_i,n}(x_1, \ldots, x_n)$. By Birg'e's Theorem, there exists a subset $\Theta_n^*$ of $\Theta_n$ such that $||\theta_1 - \theta_2||_r \geq (0.125 b_n^{-d})^{1/r} C_{q,r,a,d} b_n^{q+(d/r)}$ for all $\theta_1 \neq \theta_2$ in $\Theta_n^*$ and $\log(card(\Theta_n^*) - 1) > 0.316 b_n^{-d}$. By assumption,

$$K(P_{\theta_1(x)}, P_{\theta_2(x)}) \leq C|\theta_1(x) - \theta_2(x)|^M \leq C b_n^{Mq} \; \forall \; \theta_1, \theta_2 \in \Theta_n.$$

By Corollary 5.1, for any estimator $\hat{T}_n$,

$$\sup\{E_\theta ||\hat{T}_n - \theta||_r : \theta \in \Theta\} \geq \frac{1}{2} C_{q,r,a,d} (0.125)^{1/r} b_n^q \left[1 - \frac{C' n b_n^{Mq} + \log 2}{0.316 b_n^{-d}}\right].$$

For $[1 - \frac{C' n b_n^{Mq} + \log 2}{0.316 b_n^{-d}}]$ to be greater than a positive number, it is enough to take $b_n \sim n^{-1/(Mq+d)}$. The minimax risk cannot be better than $C^* n^{-q/(Mq+d)}$. $\square$

**Example 5.1** *In the case of conditional densities $f(y|x, \theta(x))$ that are one of the following, Bernoulli ($\theta(x)$), binomial ($N, \theta(x)$), geometric ($\theta(x)$) and exponential ($\theta(x)$), we see that $K(P_{\theta_1(x)}, P_{\theta_2(x)}) < C(\theta_1(x) - \theta_2(x))^2$ so the lower bound for the $L_r$-minimax risk is of the order $n^{-\frac{q}{2q+d}}$. The same holds for the normal $(\theta_1(x), \theta_2(x)^2)$ when we are interested in $\theta_1(x)$ and $\theta_2(x)$ is bounded away from 0 and $\infty$.*

**Example 5.2** *In the case the conditional density is either uniform ($\theta(x)$) or has the form $e^{\theta(x)-y}$, $K(P_{\theta_1(x)}, P_{\theta_2(x)}) < C|\theta_1(x) - \theta_2(x)|$ and the lower bound of the $L_r$-minimax risk is $n^{-q/(q+d)}$. One could also derive lower bounds in the case $K(P_{\theta_1(x)}, P_{\theta_2(x)}) \leq g(|\theta_1(x) - \theta_2(x)|)$.*

**Remark 5.1** *One can define d-optimality for a sequence $\hat{\theta}_n$ of estimators of $\theta$ in a regression type problem using bounds in probability for the loss $d(\hat{\theta}_n, \theta)$ (Stone, 1980). We say $\{\hat{\theta}_n\}$ is d-optimal in probability for $\theta$ if there is a sequence $\{a_n\}$ decreasing to zero such that*

$$\lim_{C \to 0} \liminf_{n \to \infty} \inf_{\hat{T}_n} \sup_{\theta \in \Theta} P[d(\hat{T}_n, \theta) > C a_n] = 1 \tag{33}$$

*and*

$$\lim_{C \to \infty} \limsup_{n \to \infty} \sup_{\theta \in \Theta} P[d(\hat{\theta}_n, \theta) > C a_n] = 0. \tag{34}$$

*To verify (33), we can use a variant of Proposition 5.1 under the additional assumption that* $d(\theta_1, \theta_2) > 2Ca_n$ *for all* $(\theta_1, \theta_2) \in \Theta_n^2, \theta_1 \neq \theta_2$. *For every estimator* $\hat{T}_n$ *we have then*

$$\sup_{\theta \in \Theta} P_\theta[d(\hat{T}_n, \theta) > Ca_n] \geq E\left[1 - \frac{(card\Theta_n)^{-2} \sum_{(\theta_1,\theta_2)\in\Theta_n^2} \sum_{i=1}^n K(P_{\theta_1(X_i)}, P_{\theta_2(X_i)}) + \log 2}{\log(card(\Theta_n) - 1)}\right],$$

*Under the previous setup, when* $\Theta$ *is the set of q-smooth functions on* $[0,1]^d, \Theta_n = \Theta_n^*, C = 0.5(0.125)^{1/r} C_{q,r,a,d}, a_n = b_n^q, K(P_{\theta_1(x)}, P_{\theta_2(x)}), \leq \tilde{C}|\theta_1(x) - \theta_2(x)|^M$ *and d is the* $L_r$ *distance,* $1 \leq r < \infty$, *we have that for every estimate* $\hat{T}_n$ *and* $\epsilon > 0, \sup_{\theta\in\Theta} P_\theta[||\hat{T}_n - \theta||_r > Cb_n^q] > 1 - \epsilon$ *for* $n \geq n(\epsilon)$ *if* $b_n \sim (\epsilon/n)^{1/(Mq+d)}$ *and*

$$\lim_{C\to 0} \liminf_{n\to\infty} \inf_{\hat{T}_n} \sup_{\theta\in\Theta} P[d(\hat{T}_n, \theta) > Cn^{-q/(Mq+d)}] = 1.$$

# 6 $L_1$-upper bound on the Error in Nonparametric Regression-Type Problems

We describe the idea for the approach in a general framework. Let $(\mathcal{X}, \mathcal{A}), (\mathcal{Y}_x, \mathcal{A}_x), x \in \mathcal{X}$ be spaces with their $\sigma$-fields and let $\mathcal{X}$ be a compact set in $R^d, d \geq 1$. $\Theta$ is a family of real-valued functions defined on $\mathcal{X}$, compact in sup-norm $||\cdot||_\infty$ on $C(\mathcal{X})$. Let $M = \{P_{x,\theta(x)}; \theta \in \Theta, x \in \mathcal{X}\}$ be a family of probability measures on $\{\mathcal{B}_x, x \in \mathcal{X}\}$ dominated by a $\sigma$-finite measure $\mu$. $P_{\theta(x)}$ will be used instead of $P_{x,\theta(x)}$ for convenience. Let $Y_1, \ldots, Y_n$ be independent r.vs, respectively, under $P_{\theta(x_i)}$ with density $f(y|x_i, \theta(x_i)), i = 1, \ldots, n$. Let $P_\theta^n$ denote the product measure $P_{\theta(x_1)} x \ldots x P_{\theta(x_n)}$ on $(\mathcal{Y}_{x_1} x \ldots x \mathcal{Y}_{x_n}, \mathcal{B}_{x_1} x \ldots x \mathcal{B}_{x_n})$. An estimate of $\theta$ is provided when the form of $P_{\theta(x)}$ is known.

Note that at each $x_i$ we have a density estimation problem. The approach in the density estimation problem suggest using the empirical measure for the solution of the regression type problem. Instead of using a MD criterion for choosing density $f_{\theta(x_i)}$ at the point $x_i$, we use a global criterionn involving densities at all the sample $x_1, \ldots, x_n$ that will allow us to choose $\theta$. Continuity of the regression function and assumption (A3) , which ensures that the observed values $x_1, \ldots, x_n$ are sufficiently dense uniformly in $\mathcal{X}$, will allow us to construct an estimate which is satisfactory globally, in $L_1$-distance.

For the regression type problem (2)

$$Y|X = x \sim P_{\theta(x)}$$

it is assumed that $\theta$ is a $q$-smooth function in $[0,1]^d$, i.e. $\theta \in \Theta_{q,d}; q = p + \alpha$. As previously discussed, $\Theta_{q,d}$ is sup-norm and totally-bounded so $L_1$ totally-bounded and has Kolmogorov entropy, $\log_2 N_\infty(a) \sim \log_2 N_{L_1}(a) \sim (\frac{1}{a})^{d/q}$.

Fundamental for the results is a Proposition showing that if $\hat{T}_n$ is an estimate of $\theta(x)$, with both $\hat{T}_n$ and $\theta$ elements of $\Theta_{q,d}$, then it is easier to estimate $\theta$ than its mixed partial derivatives $\theta^{(s)}$ as upper error bounds indicate when estimating $\theta^{(s)}$ by $\hat{T}_n^{(s)}$.

**Proposition 6.1** *(Yatracos, 1989) Under the setup of the regression type problem, for $\theta \in \Theta_{q,d}$ and $\hat{T}_n$ any estimator element of $\Theta_{q,d}, d \geq 1, 1 \leq [s] \leq p$,*

$$||\hat{T}_n^{(s)} - \theta^{(s)}||_r \leq C_1 \gamma_n^{q-[s]} + C_2 \gamma_n^{-[s]} ||\hat{T}_n - \theta||_r, \ r \geq 1; \tag{35}$$

*$C_1, C_2, \gamma_n$ are all positive constants. $\gamma_n$ is the bandwidth of a kernel used to approximate $T_n$ and $\theta$ and can be chosen to minimize the upper bound in (35) when $||\hat{T}_n - \theta||_r \leq a_n$ in probability.*

**Assumptions**

The following assumptions will be made on the distributions of the variables:

(A1) $C_1|t-s| \leq ||f((\cdot|x,t) - f((\cdot|x,s)||_1 \leq C_2|t-s|$, where $0 < C_1 \leq C_2$ are constants independent of $x$.

(A2) The form of the conditional density is known.

(A3) For every $\lambda \in (0, 1/d), d \geq 1$, there exists a $c > 0$ such that

$$\lim_{n\to\infty} Q^n(C_{n,d,\lambda}) = 1,$$

where

$$C_{n,d,\lambda} = \{(X_1, \ldots, X_n) : \#\{i : |X_i - x| < n^{-\lambda}\} \geq cn^{1-\lambda d} \text{ for all } x \in [0,1]^d\};$$

$Q^n$ is the distribution of $\mathbf{X} = (X_1, \ldots, X_n)$.

(A4) $KL(P_{\theta_1(x)}, P_{\theta_2(x)}) \leq C[\theta_1(x) - \theta_2(x)]^2$, for every $\theta_1, \theta_2$ in $\Theta_{q,d}, x \in [0,1]^d$; $C$ is a positive constant.

Assumption (A1) is satisfied in the examples that follow. You can prove it, *e.g.* for the normal and the binomial examples. Assumption (A2) is not needed when $\theta$ is the conditional mean. When data is available $f$ can be estimated locally, at each $x$, and its functional form can be approximated. Assumption (A3) is nonvacuous and has been used before in the literature; see, for example, Stone [(1982), Condition 3, page 1043]. Assumption (A4) holds under the assumptions in Proposition 5.2 and *e.g.* in Example 5.1.

Assumption (A1) is satisfied in the examples that follow. You can prove it, *e.g.* for the normal and the binomial examples.

**Example 6.1** *Normal model. Let*

$$f(y|x,\theta(x),\sigma(x)) = (2\pi)^{-1/2}(\sigma(x))^{-1}\exp\{-(y-\theta(x))^2/2\sigma^2(x)\},$$

*where $\mu$ is the Lebesgue measure. If we are interested in $\theta(x)$ and $\sigma(x)$ is bounded away from 0 and infinity on then*

$$||f(\cdot|x,\theta(x),\ \sigma(x)) - f(\cdot|x,0,\ \sigma(x))||_1 \sim |\theta(x)|,$$

*where the elements of $\Theta$ take values in $[-a,a]$ for all $x$. If we are interested in the standard deviation and if the elements of $\Theta$ (i.e., the standard deviations) are bounded away from 0 and infinity uniformly for all $x$, then*

$$||f(\cdot|x,\theta(x),\ \sigma(x)) - f(\cdot|x,0,\theta(x),\tilde{\sigma}(x))||_1 \sim |\sigma(x)-\tilde{\sigma}(x))||_1.$$

*If $\sigma(x)$ is known, then the model fits into the framework with functional parameter $\theta(x)$ and similarly for $\sigma(x)$, if $\theta(x)$ is known.*

**Example 6.2** *Exponential model. Let $f(y|t) = te^{-ty}$ where $\mu$ is the Lebesque measure on $[0,\infty)$ and $t \in [a,b] \subset (0,\infty)$, the elements of $\Theta$ taking values in $[a,b]$ for all $x \in \mathcal{X}$.*

**Example 6.3** *Poisson model. Let $f(y|t) = t^y e^{-t}/y!$ where $\mu$ is the counting measure on the nonnegative integers, $t \in [a,b] \subset (0,\infty)$.*

**Example 6.4** *Binomial model. Let $f(y|t) = \binom{N}{y}t^y(1-t)^{N-y}$ where $\mu$ is the counting measure on the nonnegative integers, $t \in [a,b] \subset (0,1)$, the elements of $\Theta$ taking values in $[a,b]$ for all $x \in \mathcal{X}$.*

**Example 6.5** *Uniform* $(0,\theta)$ *model. Let* $f(y|t) = t^{-1}, 0 \leq y \leq t$, *where* $\mu$ *is the Lebesque measure on* $[0,\infty)$ *and* $t \in [a,b] \subset (0,\infty)$, *the elements of* $\Theta$ *taking values in* $[a,b]$ *for all* $x \in \mathcal{X}$.

**Example 6.6** *Let* $f(y|t) = e^{t-y}, t \leq y$, *where* $\mu$ *is the Lebesque measure on* $[0,\infty)$ *and* $t \in [a,b] \subset (0,\infty)$, *the elements of* $\Theta$ *taking values in* $[a,b]$ *for all* $x \in \mathcal{X}$.

**Proposition 6.2** *Let* $(X_1,Y_1),\ldots,(X_n,Y_n)$ *be a sample as in the setup of the regression type problem.* $Q^n$ *is the distribution of* $\mathbf{X} = (X_1,\ldots,X_n), \theta$ *is an element of* $\Theta_{q,d}, d \geq 1$, *and* $Y_i|X_i = x_i \sim P_{\theta(x_i)}$. *Assume (A1)-(A4) hold. Then we can construct* $L_1$*-optimal minimum distance estimators* $\hat{\theta}_n$ *of* $\theta$ *with upper rate of convergence in Probability*

$$n^{-q/(2q+d)} = a_n \sim [\frac{log_2 N(a_n)}{n}]^{1/2}; \tag{36}$$

$N(a)$ *is the smallest number of* $||\cdot||_\infty$*-balls of radius* $a > 0$ *needed to cover* $\Theta_{q,d}$.

**Proof:** Fix $a_n > O$ (to be determined later in order to obtain the best convergence rate). Let $\Theta_{n,q,d}$ be the most economical $a_n$-$||\cdot||_\infty$-dense subset of $\Theta_{q,d}$ with elements $\theta_j, j = 1,\ldots,N(a_n)$.

Given that $X_1 = x_1,\ldots,X_n = x_n$,let

$$A_{k,l,i} = \{y : \frac{dP_{\theta_k(x_i)}}{d\mu}(y) > \frac{dP_{\theta_l(x_i)}}{d\mu}(y)\}, i = 1,\ldots,n, 1 \leq k < l \leq N(a_n). \tag{37}$$

The Minimum Distance estimate $\hat{\theta}_n$ is defined by

$$\sup\left\{\frac{1}{n}\left|\sum_{i=1}^{N(a_n)}\left[I_{A_k,l,i}(Y_i) - P_{\hat{\theta}_n(x_i)}(A_{k,l,i})\right]\right|; 1 \leq k < l \leq N(a_n)\right\} \tag{38}$$

$$= \inf\left\{\sup\left\{\frac{1}{n}\left|\sum_{i=1}^{N(a_n)}\left[I_{A_k,l,i}(Y_i) - P_{\theta_m(x_i)}(A_{k,l,i})\right]\right|; 1 \leq k < l \leq N(a_n)\right\}; 1 \leq m \leq N(a_n)\right\}.$$

An upper bound will be provided for $\int|\hat{\theta}_n(x) - \theta(x)|dx$ on the event $\mathbf{X} = \mathbf{x} \in C_{n,d,\lambda}$ defined in (A3). $[0,1]^d$ is covered with cubes $S_i, i = 1,\ldots,b_n^{-d}$, of side length $b_n$. Let $N_i$ be the number of coordinates $\mathbf{x} = (x_1,\ldots,x_n)$ in $S_i, M = \min\{N_i; 1 \leq i \leq b_n^{-d}\}$.

Restricting attention to $S_i$ we will approximate $\hat{\theta}_n(x)$ (resp. $\theta(x)$) with the Taylor polynomial $\hat{\theta}_{n,p}(x;x_j)$ (resp. $\theta_p(x;x_j)$) of order $p$ around each $x_j \in S_i$. From the integral form of the remainder term in Taylor's theorem, the assumption that the cube $S_i$ has side length $b_n$ and Holder's condition that holds for $\hat{\theta}_n^{(p)}(x)$ (resp. $\theta^{(p)}(x)$) one can see that the remainder is bounded in absolute value by $Cb_n^q$ in both cases; C is a positive generic constant independent of x. From now on all constants will be denoted by C. Thus we have on $S_i$,

$$\int_{S_i} |\hat{\theta}_n(x) - \theta(x)|dx \leq 2Cb_n^{q+d} + b_n^d |\hat{\theta}_n(x_j) - \theta(x_j)|dx + \sum_{1\leq[s]\leq p} b_n^{[s]} \int_{S_i} |\hat{\theta}_n^{(s)}(x_j) - \theta^{(s)}(x_j)|dx. \tag{39}$$

We now consider Taylor expansions of $\hat{\theta}_n^{(s)}(x_j), \theta^{(s)}(x_j), 1 \leq [s] \leq p$, in (39) around each $x$ in $S_i$. Proposition 6.1 is used to bound the last term of (39). For any $s,\ 1 \leq [s] \leq p$,we have, for the terms of the sum in (39),

$$b_n^{[s]} \int_{S_i} |\hat{\theta}_n^{(s)}(x_j) - \theta^{(s)}(x_j)|\ dx \leq b_n^{[s]} \left[ 2Cb_n^{q+d-[s]} + \int_{S_i} |\hat{\theta}_{n,p-[s]}^{(s)}(x_j;x) - \theta_{p-[s]}^{(s)}(x_j;x)|\ dx \right]$$

$$\leq 2Cb_n^{q+d} + \sum_{0\leq[t]\leq p-[s]} b_n^{s+t} \int_{S_i} |\hat{\theta}_n^{(s+t)}(x) - \theta^{(s+t)}(x)|\ dx, \tag{40}$$

where $s+t$ is the usual sum between vectors.

From (39) and (40) we obtain

$$\int_{S_i} |\hat{\theta}_n(x) - \theta(x)|dx \leq Cb_n^{q+d} + b_n^d |\hat{\theta}_n(x_j) - \theta(x_j)| + C \sum_{1\leq[s]\leq p} \sum_{0\leq[t]\leq p-[s]} b_n^{s+t} \int_{S_i} |\hat{\theta}_n^{(s+t)}(x) - \theta^{(s+t)}(x)|\ dx$$

$$\leq Cb_n^{q+d} + b_n^d |\hat{\theta}_n(x_j) - \theta(x_j)| + C \sum_{1\leq[s]\leq p} b_n^{[s]} \int_{S_i} |\hat{\theta}_n^{(s)}(x) - \theta^{(s)}(x)|\ dx. \tag{41}$$

Repeating (41) for $M$ out of the $N_i$ elements in $S_i$ and for all $i$ we obtain

$$M||\hat{\theta}_n - \theta||_1 \leq CMb_n^q + b_n^d \sum_{j=1}^{n} |\hat{\theta}_n(x_j) - \theta(x_j)| + CM \sum_{1\leq[s]\leq p} b_n^{[s]} ||\hat{\theta}_n^{(s)} - \theta^{(s)}||_1. \tag{42}$$

From $(A3)$ on the event $C_{n,d,i}$ with $b_n = n^{-\lambda}$, one has $cn^{1-\lambda d} \leq N_i$ for all $i$, and

$$cnb_n^d = cn^{1-\lambda d} \leq M = \min\{N_i; 1 \leq i \leq b_n^{-d}\}.$$

Thus,

$$||\hat{\theta}_n - \theta||_1 \leq Cb_n^q + n^{-1}\sum_{j=1}^{n}|\hat{\theta}_n(x_j) - \theta(x_j)| + C\sum_{1\leq[s]\leq p} b_n^{[s]}||\hat{\theta}_n^{(s)} - \theta^{(s)}||_1. \tag{43}$$

From assumption (A1) and the definition of $\hat{\theta}_n$ as minimum distance estimate via $\Theta_{n,q,d}$ we obtain

$$n^{-1}\sum_{j=1}^{n}|\hat{\theta}_n(x_j)-\theta(x_j)| \leq Ca_n+Cn^{-1}\sup\left\{\left|\sum_{j=1}^{n}[P_{\theta(x_j)}(A_{k,l,j}) - I_{A_{k,l,j}}(Y_j)]\right| 1\leq k<l\leq N(a_n)\right\}. \tag{44}$$

From (43), (44) and Proposition 6.1 with $\gamma_n = D\cdot b_n,\ D$ a positive constant large enough, we obtain

$$||\hat{\theta}_n-\theta||_1(1-\frac{C}{D}) \leq Cb_n^q+Ca_n+Cn^{-1}\sup\left\{\left|\sum_{j=1}^{n}[P_{\theta(x_j)}(A_{k,l,j}) - I_{A_{k,l,j}}(Y_j)]\right| 1\leq k<l\leq N(a_n)\right\}. \tag{45}$$

Note that in (45) it holds $(1-\frac{C}{D})>0$, since $D$ can be chosen as large as we wish.

A bound in probability can be derived for the random variable in the right side of (45) as we have done previously to obtain with probability tending to one for $\mathbf{x}\in C_{n,d,\lambda}$ that

$$||\hat{\theta}_n - \theta||_1 \leq C\left[a_n + b_n^q + \left(\frac{\log_2 N(a_n)}{n}\right)^{1/2}\right]. \tag{46}$$

Thus, given $\mathbf{x}\in C_{n,d,\lambda}$, an upper bound in (46) is obtained by choosing $a_n$ and $b_n$ such that $a_n = b_n^q = n^{-q/(2q+d)}, a_n \sim (\frac{\log_2 N(a_n)}{n})^{1/2}$. Note that $\lambda = (q+2d)^{-1} < d^{-1}$ as required in (A3).

Finally,

$$\begin{aligned}P[||\hat{\theta}_n - \theta||_1 > Ca_n] &= E_{Q^n}P\left[||\hat{\theta}_n - \theta||_1 > Ca_n|\mathbf{X}=\mathbf{x}\right] I(\mathbf{x}\in C_{n,d,1/(2q+d)}) \\ &+E_{Q^n}P\left[||\hat{\theta}_n - \theta||_1 > Ca_n|\mathbf{X}=\mathbf{x}\right] I(\mathbf{x}\in C^c_{n,d,1/(2q+d)}) \to 0\end{aligned}$$

as $n$ increases to infinity by means of (46) and (A3).

This result holds also when $p=0$. Using the Lipschitz condition we obtain inequalities (39), (42), and (43) without the terms that involve derivatives; (45) holds with $(1-C/D)$ replaced by 1. The rest follows as in the case where $p>0$. Using assumption (A4), the estimate $\hat{\theta}_n$ is optimal in Example 5.1 for which

$$K(P_{\theta_1(x)}, P_{\theta_2(x)}) \leq C[\theta_1(x)-\theta_2(x)]^2.$$

Kernel estimates of a density $f$ [1]

Let $X, X_1, \ldots, X_n$ be i.i.d. random variables, with density $f$ and cumulative distribution function $F$, $F'(x) = f(x)$. Here there is no unknown model parameter and $f$ is unknown. This is a nonparametric estimation problem. The goal is to estimate $f(x)$. Recall that the empirical distribution function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i),$$

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h} \approx \frac{F(x+h) - F(x-h)}{2h} \text{ for } h \text{ small.}$$

Thus, an estimate of $f$ is

$$\hat{f}_{n,U}(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2nh} = \frac{1}{nh}\sum_{i=1}^{n} \frac{I(|\frac{X_i - x}{h}| \le 1)}{2} = \frac{1}{nh}\sum_{i=1}^{n} K_U(\frac{X_i - x}{h})$$

with $K_U(y)$ a uniform density on $(-1, 1)$, *i.e.* $K_U(y) = 1/2$ if $|y| \le 1$ and vanishes otherwise. The form of $\hat{f}_{n,U}$ motivates its generalization using any density $K$ :

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h}),\ h = h_n \text{ (to be determined).} \tag{47}$$

Observe that

$$\int_{-\infty}^{+\infty} \frac{1}{h} K(\frac{X_i - x}{h})dx = \int_{-\infty}^{+\infty} K(y)dy = 1 \to \int_{-\infty}^{+\infty} \hat{f}_n(x) = 1,$$

thus $\hat{f}_n$ is a density when $K$ is a density.

Question: Note that $\hat{f}_n(x) = \hat{f}_n(x; X_1, \ldots, X_n)$. Using density estimate $\hat{f}_n(x; X_1, \ldots, X_n)$, calculate the estimates of the mean, $EX$, and the variance, $E(X^2) - (EX)^2$, which will be both functions of $X_1, \ldots, X_n$ and compare them, respectively, with the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the sample variance, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

**Definition 6.1** *A function $K : R \to R$ is called Kernel when $\int_{-\infty}^{+\infty} K(y)dy = 1$.*

[1] From B. Hansen's "Lecture Notes on Nonparametrics" in the Web, https://www.ssc.wisc.edu/ bhansen/718/NonParametrics1.pdf .

A non-negative Kernel $K$ is a density and then $\hat{f}_n$ is a density. The $j$-th moment of Kernel $K$ is

$$m_j = \int_{-\infty}^{+\infty} y^j K(y)dy. \tag{48}$$

For a symmetric around 0 Kernel $K$ the odd moments $m_{2j+1}$ are zero, $j = 0, 1, \ldots$. Usually, in nonparametric estimation symmetric Kernels $K$ are used. The order $\nu$ of Kernel $K$ is the order of its first non-zero moment.

Higher order Kernel: any kernel for which $\nu > 2$.

Examples of Kernels

Epanechnikov kernel: $K_E(y) = \frac{3}{4}(1-y)^2, |y| \leq 1$ and 0 otherwise.
Gaussian kernel: $K_G(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \phi(y),\ -\infty < y < \infty$.

What are the orders $\nu$ of $K_E, K_G$?

Kernels with order $\nu$ higher than 2 can be obtained from second order Kernels multiplied with a polynomial in $y^2$ of degree $\frac{\nu}{2} - 1$.

Construction for $\nu = 4$-th order Kernel with $K_G$: Since $\nu = 4 \to (\frac{\nu}{2} - 1) = 1$ the new Kernel will have form

$$K_{NEW}(y) = \phi(y)(a + by^2).$$

Need two conditions to determine $a, b$ :

$$1 = \int_{-\infty}^{+\infty} K_{NEW}(y)dy = a + b$$

$$0 = \int_{-\infty}^{+\infty} y^2 K_{NEW}(y)dy = a+b\int_{-\infty}^{+\infty} y^4\phi(y)dy = a-b\int_{-\infty}^{+\infty} y^3 d\phi(y) = a+3b\int_{-\infty}^{+\infty} y^2\phi(y)dy = a+3b.$$

Thus, $b = \frac{-1}{2},\ a = \frac{3}{2}$ and $K_{NEW}(y) = \frac{1}{2}(3-u^2)\phi(u),\ u \in R$.

Practice: Show that *a)* a 4-th order Kernel for $K_E$ is $\frac{15}{8}(1 - \frac{7}{3}y^3)K_E(y), -1 \leq y \leq 1$ and *b)* a 6-th order kernel for $K_G$ is $\frac{1}{8}(15 - 10y^2 + y^4)K_G(y),\ u \in R$.

Investigating local at $x$ properties of the estimate $\hat{f}_n(x; X_1, \ldots, X_n)$ of $f(x)$

Bias of $\hat{f}_n(x; X_1, \ldots, X_n)$

$$E\hat{f}_n(x; X_1, \ldots, X_n) = E\frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h}) = \frac{1}{h}EK(\frac{X_1 - x}{h}) = \int_{-\infty}^{+\infty} \frac{1}{h}K(\frac{y-x}{h})f(y)dy$$

and with the transformation $u = \frac{y-x}{h} \rightarrow du = dy/h, y = x + uh$

$$E\hat{f}_n(x; X_1, \ldots, X_n) = \frac{1}{h} EK(\frac{X_1 - x}{h}) = \int_{-\infty}^{+\infty} f(x + uh)K(u)du.$$

Since $f$ is unknown, Taylor's Theorem with Remainder will be used assuming $f \in \mathcal{C}^{k+1}$, *i.e.* has $(k+1)$-continuous derivatives, obtaining the approximation:

$$f(x + hu) = f(x) + huf'(x) + \frac{h^2u^2}{2!} f^{(2)}(x) + \ldots + \frac{h^k u^k}{k!} f^{k)}(x) + o(h^k). \tag{49}$$

Therefore, the bias of $\hat{f}_n(x; X_1, \ldots, X_n)$ in estimating $f(x)$ is

$$E\hat{f}_n(x; X_1, \ldots, X_n) - f(x) = hm_1 f'(x) + \frac{h^2}{2!} m_2 f^{(2)}(x) + \ldots + \frac{h^k}{k!} f^{k)}(x) m_k + o(h^k). \tag{50}$$

**Remark 6.1** *For symmetric kernels with $\nu = 2$, the order of the bias is $\frac{h^2}{2!} m_2 f^{(2)}(x)$ but for higher order kernels, e.g. $\nu = k(even)$ the order of the bias is proportional to $h^k$, i.e. reduced.*

$h$ is called <u>bandwidth</u> and observe that, as expected, the smaller $h$ is the smaller the bias in (50) is.

<u>Variance of $\hat{f}_n(x; X_1, \ldots, X_n)$</u>

Since $X_1, \ldots, X_n$ are i.i.d.

$$Var(\hat{f}_n(x; X_1, \ldots, X_n)) = Var(\frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h})) = \frac{1}{nh^2} Var(K(\frac{X_1 - x}{h}))$$

$$= \frac{1}{nh^2}[E(K(\frac{X_1 - x}{h})^2) - (EK(\frac{X_1 - x}{h}))^2] = \frac{1}{nh^2} E[K(\frac{X_1 - x}{h})^2] - \frac{1}{n}[\frac{1}{h} EK(\frac{X_1 - x}{h})]^2$$

From (49),

$$\frac{1}{n} E\frac{1}{h} K(\frac{X_1 - x}{h}) = \frac{1}{n}[f(x) + o(1)] = O(\frac{1}{n}),$$

and

$$\frac{1}{nh^2} E[K(\frac{X_1 - x}{h})^2] = \frac{1}{nh} \int_{-\infty}^{+\infty} K^2(y) f(x+yh)dy = \frac{1}{nh} \int_{-\infty}^{+\infty} K^2(y)[f(x)+o(h)]dy = \frac{f(x) \int_{-\infty}^{+\infty} K^2(y)dy}{nh}.$$

Therefore,

$$Var(\hat{f}_n(x; X_1, \ldots, X_n)) \approx \frac{f(x) \int_{-\infty}^{+\infty} K^2(y)dy}{nh}. \tag{51}$$

**Remark 6.2** *$Var(\hat{f}_n(x; X_1, \ldots, X_n))$ increases as $h$ decreases, in the opposite direction of the bias.*

**Remark 6.3** *Observe that $Var(\hat{f}_n)$ is not affected by the order of the kernel $K$ unlike its bias.*

Mean Square Error (MSE) of $\hat{f}_n(x; X_1, \ldots, X_n)$

$$MSE(\hat{f}_n(x), f(x)) = E[\hat{f}_n(x) - f(x)]^2 = Var(\hat{f}_n(x)) + Bias^2(\hat{f}_n(x))$$

It follows from (50) and (51) that for a kernel with order $\nu = k$,

$$MSE(\hat{f}_n(x), f(x)) \approx \frac{f(x)\int_{-\infty}^{+\infty} K^2(y)dy}{nh} + h^{2\nu}[\frac{f^{(\nu)}(x)m_\nu}{\nu!}]^2 = AMSE(\hat{f}_n(x), f(x)). \quad (52)$$

Approximation (52) is a local at $x$ asymptotic result in $n$ with $h = h_n$ to be determined, and this is why it is called Asymptotic MSE (AMSE). From (52), for the AMSE to decrease to zero as $n$ increases it should hold:

$$\lim_{n\to\infty} h_n = 0 \qquad \text{and} \qquad \lim_{n\to\infty} n \cdot h_n = \infty. \quad (53)$$

Then, the interval $(x - h_n, x + h_n)$ used will be small enough for the bias to be controlled (and decrease), but also large enough such that the proportion of the $n$ observations it contains, which is approximately $n \cdot f(x) \cdot 2h_n$, will increase to infinity for the variance to decrease to zero.

To derive an optimal value of $h_n = h_n(x)$ that minimizes the $AMSE(\hat{f}_n(x), f(x))$ solve the equation

$$0 = \frac{d}{dh} AMSE(\hat{f}_n(x), f(x)) = -\frac{f(x)\int_{-\infty}^{+\infty} K^2(y)dy}{nh^2} + 2\nu h^{2\nu-1}[\frac{f^{(\nu)}(x)m_\nu}{\nu!}]^2$$

$$= -\frac{C_1(f(x), K)}{nh^2} + C_2(f^{(\nu)}(x), m_\nu, \nu)h^{2\nu-1} \to h = h_n = \frac{C_1(f(x), K)}{C_2(f^{(\nu)}(x), m_\nu, \nu)n^{1/2\nu+1}},$$

which minimizes $AMSE(\hat{f}_n(x), f(x))$ since $C_1(f(x), K) > 0,\ C_2(f^{(\nu)}(x), m_\nu, \nu) > 0$. The constants of proportionality are usually unknown. Thus, the optimal bandwidth for the $AMSE$ is $h_n = O(n^{-\frac{1}{2\nu+1}})$; this confirms that higher order kernels can afford larger bandwidth. The optimal $AMSE(\hat{f}_n(x), f(x)) = O(n^{-\frac{2\nu}{2\nu+1}})$.

Global error for all $x$

One could look either at $\sup_{x\in R} AMSE(\hat{f}_n(x), f(x))$ or at $\int_R AMSE(\hat{f}_n(x), f(x))dx$. The latter is often called Asymptotic Mean Integrated Square Error,

$$AMISE(\hat{f}_n, f) = \int_R AMSE(\hat{f}_n(x), f(x))dx = \frac{\int_{-\infty}^{+\infty} K^2(y)dy}{nh} + h^{2\nu}\frac{m_\nu^2}{\nu!^2}\int_R [f^{(\nu)}(x)]^2 dx. \tag{54}$$

For the optimal bandwidth the steps used for $AMSE(\hat{f}_n(x), f(x))$ are followed and the results are similar, with the coefficients in the $Big - O$ notation changed and

$$AMISE_{opt}(\hat{f}_n, f)(K) = (1+2\nu)[\frac{(\int_R K^2(y)dy)^{2\nu} \cdot m_\nu^2(K) \cdot \int_R (f^{(\nu)}(x))^2 dx}{(\nu!)^2(2\nu)^{2\nu}}]^{1/(2\nu+1)} \cdot n^{-2\nu/(2\nu+1)} \tag{55}$$

Observe that for a second-order kernel, i.e. with $\nu = 2$, $AMISE_{opt}(\hat{f}_n, f) = O(n^{-4/5})$. Making the $MSE$ distance by taking its square root, the rate is $\sqrt{AMISE_{opt}(\hat{f}_n, f)} = O(n^{-2/5})$. We have already seen rates of convergence of estimates with respect to other distances. Observe also that since

$$\sqrt{AMISE_{opt}(\hat{f}_n, f)(K)} \leq Cn^{-\nu/(2\nu+1)}$$

when $\nu \to \infty$ the order of the upper bound converges to $n^{-1/2}$ which is the parametric rate of convergence for estimates.

Comparison of Kernel estimates

Given order $\nu$ Kernels $K_1$, $K_2$, one could compare them by looking at the ratio of their $AMISE_{opt}$. For estimation of the density function, the higher-order Epanechikov kernel with optimal bandwidth yields the lowest possible $AMISE$. For this reason, the Epanechikov kernel is often called the "optimal kernel".

More on density estimates in the Lecture Notes by B. Hansen and the references there in as well as the works of Devroye and his co-authors.

# References

[1]

[2]