

April 23, 2003

# NONPARAMETRIC LOGISTIC REGRESSION: REPRODUCING KERNEL HILBERT SPACES AND STRONG CONVEXITY

P.P.B. EGGERMONT and V.N. LARICCIA

University of Delaware

ABSTRACT. We study maximum penalized likelihood estimation for logistic regression type problems. The usual difficulties encountered when the log-odds ratios may become large in absolute value are circumvented by imposing a priori bounds on the estimator, depending on the sample size ( $n$ ) and smoothing parameter. We pay for this in the convergence rate of the mean integrated squared error by a factor  $\log n$ . When the “true” log odds ratios are bounded, these factors  $\log n$  disappear. The main technical tools employed are reproducing kernel Hilbert spaces and convexity inequalities.

## 1. INTRODUCTION

We are interested in the estimation of smooth functions by means of the maximum penalized likelihood approach. In density as well as regression function estimation, this approach works quite well but proving acceptable convergence rates under acceptable conditions turns out to be quite hard. Of course, the standard example of penalized least-squares regression estimation (a.k.a. spline smoothing) may be handled beautifully by reproducing kernel Hilbert space methods, see, e.g., EUBANK (1999). Indeed, for “general” maximum penalized likelihood estimation, Newton’s method, when applicable, reduces the problem to penalized weighted least-squares, see, e.g., COX and O’SULLIVAN (1990), KOOPERBERG and STONE (1991), GU and QIU (1993) and GU (2002). If one does not want to get involved with the quadratic approximations and the assumptions required to make it work, some fundamental problems, having to do with tail behavior, must be resolved. Successful treatments then seem to depend on lucky accidents, see, e.g., the authors’ paper (1999) and book

(2001, pp. 189–201). Here, we investigate what can be done in more general situations.

As our vehicle, we consider nonparametric logistic regression, misnomer as that may be. We hasten to add that logistic regression is important in its own right, see, e.g., KLEINBAUM and KLEIN (2002) and references therein. Let

$$(1.1) \quad 0 \leq x_{1,n} < x_{2,n} < \cdots < x_{nn} \leq 1$$

be an asymptotically uniform deterministic design on  $[0, 1]$  (see §2 for the precise meaning of asymptotic uniformity) and suppose we observe the independent random variables  $y_{1,n}, y_{2,n}, \dots, y_{nn}$ , with

$$(1.2) \quad \mathbb{P}[y_{in} = 1] = p_o(x_{in}) \quad , \quad \mathbb{P}[y_{in} = 0] = 1 - p_o(x_{in})$$

for  $i = 1, 2, \dots, n$ . Here,  $p_o$  is a nice, smooth function on  $[0, 1]$ . Situations like these are not uncommon in bioassay studies, see, e.g., FINNEY (1992).

For the model (1.2), the negative log-likelihood of any feasible  $p$  is

$$(1.3) \quad L_n(p) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \{ y_{in} \log p(x_{in}) + (1 - y_{in}) \log(1 - p(x_{in})) \} ,$$

and *maximum penalized likelihood estimation* takes the form

$$(1.4) \quad \begin{array}{ll} \text{minimize} & L_n(p) + h^2 \text{RP}(p) \\ \text{subject to} & 0 \leq p \leq 1 \text{ everywhere} , \end{array}$$

where  $\text{RP}(p)$  is the roughness penalization. The choice

$$(1.5) \quad \text{RP}(p) = \int_0^1 \frac{|p'(t)|^2}{p(t)(1-p(t))} dt$$

stands out: Writing

$$(1.6) \quad \frac{|p'(t)|^2}{p(t)(1-p(t))} = \frac{|p'(t)|^2}{p(t)} + \frac{|(1-p(t))'|^2}{1-p(t)}$$

shows that  $\text{RP}(p)$  is the Fisher information for the nonparametric logistic regression problem (1.4). For related observations see, e.g., KASS (1989). The roughness penalization under consideration is quite closely related to the roughness penalization of GOOD (1971) for nonparametric density estimation and, indeed, the Bayesian arguments of GOOD (1971) apply here as well.

The problem (1.4) being strictly convex, it is a standard exercise to show that its solution exists and is unique; we denote it by  $p = \pi^{nh}$ . Now, our goal is to prove convergence rates for the estimator  $\pi^{nh}$ . Without straightaway going into the many details, it seems clear that this involves establishing convergence rates for  $L_n(p_o) - L_n(\pi^{nh}) \rightarrow 0$ , or

$$(1.7) \quad \frac{1}{n} \sum_{i=1}^n \left\{ y_{in} \log \frac{\pi^{nh}}{p_o(x_{in})} + (1 - y_{in}) \log \frac{1 - \pi^{nh}}{1 - p_o(x_{in})} \right\} \rightarrow 0$$

in a suitable sense. One is thus faced with the task of controlling the logarithms of the ratios in question, especially when  $p_o$  is either close to zero or one, and this the authors are unable to do. As a way out, we impose bounds on these ratios as part of the minimization problem: We propose to estimate  $p_o$  by the solution  $p^{nh}$  to

$$(1.8) \quad \begin{aligned} & \text{minimize} && L_n(p) + h^2 \text{RP}(p) \\ & \text{subject to} && \text{RP}(p) < \infty, \\ & && (nh)^{-1} \leq p \leq 1 - (nh)^{-1}. \end{aligned}$$

The reason for this particular choice of the bounds will become clear later. One should keep in mind that this is a theoretical proposal: Now, we are able to prove convergence rates for a rational estimator. Simulation experiments will have to decide whether the estimator defined by (1.4) is to be preferred in practice.

Some comments are in order. First, if  $p_o$  is bounded away from zero and one, then the device (1.8) is not necessary. See Remark (1.20) below. Second, the problem (1.8) is again/still strictly convex and its solution exists and is unique. We denote it by  $p^{nh}$ . Third, it is rather annoying that  $p_o$  need not satisfy the constraints of (1.8). This can be fixed by shrinking towards the “middle”, i.e., by replacing  $p_o$  by  $p_h$  defined as

$$(1.9) \quad p_h(t) = \frac{p_o(t) + \lambda}{1 + 2\lambda}, \quad t \in [0, 1]$$

with  $\lambda = \{nh - 2\}^{-1}$ . Note that  $\lambda/(1 + 2\lambda) = 1/(nh)$ .

Then, obviously,

$$(1.10) \quad \left| \log \frac{p^{nh}(t)}{p_h(t)} \right| \leq \log(nh - 1) \leq \log nh ,$$

and likewise for the other ratio. Of course, in (1.8) and (1.10) it is required that  $nh > 2$ , but  $nh \rightarrow \infty$  is a typical requirement for the convergence of nonparametric function estimators.

In the remainder of this paper, we study convergence rates for the estimator  $p^{nh}$ . First, for nice functions  $p$  and  $\pi$  on  $[0, 1]$ , define the Kullback-Leibler distance

$$(1.11) \quad \text{KL}(p, \pi) = \int_0^1 \left\{ p(t) \log \frac{p(t)}{\pi(t)} + \pi(t) - p(t) \right\} dt ,$$

as well as its discrete analogue

$$(1.12) \quad \text{KL}_n(p, \pi) = \frac{1}{n} \sum_{i=1}^n \left\{ p(x_{in}) \log \frac{p(x_{in})}{\pi(x_{in})} + \pi(x_{in}) - p(x_{in}) \right\} .$$

Secondly, for  $0 \leq p, \pi \leq 1$ , define the ‘‘logistic distance’’, the sum of the Kullback-Leibler distances between  $p$  and  $\pi$  and between  $1 - p$  and  $1 - \pi$ , and its discrete analogue

$$(1.13) \quad \begin{aligned} \Lambda(p, \pi) &= \text{KL}(p, \pi) + \text{KL}(1 - p, 1 - \pi) , \\ \Lambda_n(p, \pi) &= \text{KL}_n(p, \pi) + \text{KL}_n(1 - p, 1 - \pi) . \end{aligned}$$

In order to establish convergence rates, some assumptions must be made. A minimal one is that

$$(1.14) \quad \text{RP}(p_o) < \infty ,$$

and a good case can be made for the boundary conditions

$$(1.15) \quad p'_o(0) = p'_o(1) = 0 .$$

We shall even assume that

$$(1.16) \quad \left\| \left\{ \sqrt{p_o} \right\}'' \right\|^2 + \left\| \left\{ \sqrt{1 - p_o} \right\}'' \right\|^2 < \infty .$$

(1.17) MAIN THEOREM. *If  $p_o$  satisfies (1.14) then the solution  $p^{nh}$  of (1.8) satisfies*

$$\mathbb{E} \left[ \Lambda(p_h, p^{nh}) \right] = \mathcal{O}(n^{-2/3} \log n) ,$$

*provided  $h \asymp n^{-1/3}$  (deterministically).*

*If, moreover,  $p_o$  satisfies (1.15)-(1.16), then the rate improves to*

$$\mathbb{E} \left[ \Lambda(p_h, p^{nh}) \right] = \mathcal{O}(n^{-4/5} \log n) ,$$

*provided  $h \asymp n^{-1/5}$ , again deterministically.*

One should note the inequality

$$(1.18) \quad \|p_h - p^{nh}\|^2 \leq \Lambda(p_h, p^{nh}) ,$$

see, e.g., KEMPERMAN (1967), as well as  $\|p_h - p_o\|^2 = \mathcal{O}(h^4 + (nh)^{-1})$ , provided (1.14) through (1.16) hold, so that the theorem implies the bound

$$(1.19) \quad \mathbb{E}[\|p^{nh} - p_o\|^2] = \mathcal{O}(n^{-4/5} \log n) .$$

Of course, for penalized least-squares estimation or local polynomial least squares, one obtains the rate  $n^{-4/5}$ , see, e.g., STONE (1982), so the conjecture is that in Theorem (1.17) and (1.19), the  $\log n$  factors are not necessary.

(1.20) REMARK. If  $p_o$  is bounded away from zero and one, then the factors  $\log n$  in the bounds proposed in Theorem (1.17) are not necessary. Proving this amounts to showing that the constrained solution of (1.8) satisfies

$$(1.21) \quad \|p^{nh} - p_o\|_\infty \longrightarrow 0 \quad \text{almost surely} ,$$

which implies that we may strengthen the constraints in (1.8) to

$$(1.22) \quad \delta \leq p(t) \leq 1 - \delta , \quad t \in [0, 1] ,$$

for an appropriate constant  $\delta$  (so that the factors  $\log n$  in Theorem (1.17) disappear) and, at the same time, that the constraints (1.22) are not active (so that the solutions of (1.8)-(1.22) and the unrestricted problem (1.4) coincide). The main details are shown in §9.

## 2. THE INGREDIENTS OF THE PROOF

We only consider part (b) of the Main Theorem (1.17). There are two fundamental ingredients in the proof, to wit a strong-convexity inequality and a reproducing kernel Hilbert space trick. Unfortunately, there are also some annoying technical details to be taken care of.

In what follows, the relevant class of functions is

$$(2.1) \quad \mathcal{D}_{nh} = \left\{ p : [0, 1] \rightarrow [0, 1] \left| \begin{array}{l} (nh)^{-1} \leq p \leq 1 - (nh)^{-1}, \\ p \text{ absolutely continuous} \\ \text{and } \text{RP}(p) < \infty \end{array} \right. \right\}.$$

On the set  $\mathcal{D}_{nh}$ , a collection of squared “distances” is needed. Besides the logistic Kullback-Leibler distance, we also need “quadratic” versions,

$$(2.2) \quad \text{QKL}(p, \pi) = \int_0^1 p(t) \left| \log \frac{p(t)}{\pi(t)} \right|^2 dt$$

and

$$(2.3) \quad \text{QA}(p, \pi) = \text{QKL}(p, \pi) + \text{QKL}(1 - p, 1 - \pi),$$

together with their discrete analogues

$$(2.4) \quad \text{QKL}_n(p, \pi) = \frac{1}{n} \sum_{i=1}^n p(x_{in}) \left| \log \frac{p(x_{in})}{\pi(x_{in})} \right|^2,$$

and

$$(2.5) \quad \text{QA}_n(p, \pi) = \text{QKL}_n(p, \pi) + \text{QKL}_n(1 - p, 1 - \pi).$$

Finally, differences between derivatives are measured by means of

$$(2.6) \quad \Delta(p, \pi) = \int_0^1 p(t) \left| \left\{ \log \frac{p(t)}{\pi(t)} \right\}' \right|^2 dt$$

and

$$(2.7) \quad \Pi(p, \pi) = \Delta(p, \pi) + \Delta(1 - p, 1 - \pi).$$

**The first fundamental step** in the proof of the Main Theorem is the strong-convexity-like inequality

$$(2.8) \quad \Lambda_n(p_h, p^{nh}) + h^2 \Pi(p_h, p^{nh}) \leq S_n(p_h, p^{nh}) + h^2 \left\{ \text{RP}(p_h) - \text{RP}(p^{nh}) \right\},$$

where

$$(2.9) \quad S_n(p, \pi) = \frac{1}{n} \sum_{i=1}^n \left\{ y_{in} - p(x_{in}) \right\} \left\{ \log \frac{p(x_{in})}{\pi(x_{in})} - \log \frac{1 - p(x_{in})}{1 - \pi(x_{in})} \right\}.$$

The inequality (2.8) also holds with  $p_h$  replaced by  $p_o$ , but this leads to the problems outlined in the introduction. The standard but somewhat elaborate proof of (2.8) is given in § 3.

**The second fundamental step** consists of using reproducing kernel Hilbert space (rkhs) methods. In its simplest form, the rkhs method shows that for  $\varepsilon_{1,n}, \varepsilon_{2,n}, \dots, \varepsilon_{n,n}$  uncorrelated, zero mean random variables with bounded variance, and random  $f$  with square integrable derivative,

$$(2.10) \quad \left\{ \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} f(x_{in}) \right] \right\}^2 \leq c (nh)^{-1} \mathbb{E} \left[ \|f\|^2 + h^2 \|f'\|^2 \right].$$

A suitable modification of the argument leads to the precise bound

$$(2.11) \quad \left\{ \mathbb{E} [S_n(p_h, p^{nh})] \right\}^2 \leq c h^4 (nh)^{-1} (\log nh)^2 + c (nh)^{-1} \mathbb{E} \left[ \text{QA}(p_h, p^{nh}) + h^2 \Pi(p_h, p^{nh}) \right].$$

The first term on the right is one of the messy penalties we pay for the truncation (1.8). All of this is shown in §§ 4 and 5.

After taking expectations in (2.8), using the bound (2.11) yields

$$(2.12) \quad \mathbb{E} \Lambda_n + h^2 \mathbb{E} \Pi \leq c h^2 (nh)^{-1/2} \log nh + c (nh)^{-1/2} \left\{ \mathbb{E} \text{QA} + h^2 \mathbb{E} \Pi \right\}^{1/2} + h^2 \left\{ \text{RP}(p_h) - \mathbb{E} \text{RP}(p^{nh}) \right\}.$$

Here, we dropped the arguments  $(p_h, p^{nh})$  of  $\Lambda_n$  and  $\Pi$ .

**The remaining steps are mostly** annoying technical details. One may get rid of  $Q\Lambda$  by means of the elementary inequality, valid for all  $p, \pi \in \mathcal{D}_{nh}$ ,

$$(2.13) \quad Q\Lambda(p, \pi) \leq c \Lambda(p, \pi) \log nh ,$$

where  $c = c(nh) \rightarrow 1$  as  $nh \rightarrow \infty$ . See § 6.

Substituting the bound (2.13) into (2.12) reveals that we must reconcile the distinction between the discrete and continuous versions of the logistic distance,  $\Lambda_n$  and  $\Lambda$ . This difference is a quadrature error. Indeed, in § 7, we show that the uniform designs

$$(2.14) \quad x_{in} = \frac{i - \frac{1}{2}}{n - 1} \quad \text{or} \quad x_{in} = \frac{i - 1}{n - 1} \quad , \quad i = 1, 2, \dots, n ,$$

are asymptotically uniform in the sense of the following definition.

(2.15) DEFINITION. *A family of deterministic designs  $x_{in}, i = 1, 2, \dots, n$ , is asymptotically uniform if there exists a constant  $c$  such that for all deterministic, continuous functions  $f$  with integrable derivative and all  $n$ ,*

$$\left| \frac{1}{n} \sum_{i=1}^n f(x_{in}) - \int_0^1 f(t) dt \right| \leq c n^{-1} \|f'\|_1 .$$

Applied to  $\Lambda_n$ , this results in

$$(2.16) \quad \left| \Lambda_n(p, \pi) - \Lambda(p, \pi) \right| \leq c \frac{(\log nh)^{1/2}}{nh} \left\{ \Lambda(p, \pi) + h^2 \Pi(p, \pi) \right\}^{1/2}$$

for all  $p, \pi \in \mathcal{D}_{nh}$ . See § 7.

Consequently, for all  $p, \pi \in \mathcal{D}_{nh}$ , barring  $p = \pi$  everywhere,

$$(2.17) \quad 1 - \eta_{nh} \leq \frac{\Lambda_n(p, \pi) + h^2 \Pi(p, \pi)}{\Lambda(p, \pi) + h^2 \Pi(p, \pi)} \leq (1 - \eta_{nh})^{-1} ,$$

where  $\eta_{nh} = c(nh)^{-1}(\log nh)^{1/2}$  for a suitable constant depending on  $p_o$ . So,  $\eta_{nh} \rightarrow 0$  for  $nh \rightarrow \infty$ . The inequality (2.17) implies that in (2.12), we may safely ignore the difference between  $\Lambda_n$  and  $\Lambda$ .

**Finally**, in § 8, we show that

$$(2.18) \quad \mathbb{R}P(p_h) - \mathbb{E}[\mathbb{R}P(p^{nh})] \leq c \left\{ \mathbb{E}[\Lambda(p_h, p^{nh})] \right\}^{1/2}$$

provided  $p_o$  satisfies (1.14)-(1.16). Since we must substitute this into (2.12), it is easier to replace (2.18) by

$$(2.19) \quad \mathbb{R}P(p_h) - \mathbb{E}[\mathbb{R}P(p^{nh})] \leq c \left\{ \mathbb{E}[\Lambda(p_h, p^{nh}) + h^2 \Pi(p_h, p^{nh})] \right\}^{1/2} .$$

**All that remains** is substituting these bounds into (2.12). This gives, for suitable constants  $c$ ,

$$(2.20) \quad \mathbb{E}\Lambda + h^2 \mathbb{E}\Pi \leq c h^2 (nh)^{-1/2} \log nh + (\varepsilon + c h^2) (\mathbb{E}\Lambda + h^2 \mathbb{E}\Pi)^{1/2} ,$$

where  $\varepsilon^2 = c (nh)^{-1} \log nh$  and  $\Lambda \equiv \Lambda(p_h, p^{nh})$ ,  $\Pi \equiv \Pi(p_h, p^{nh})$ . Now, move the second term on the right hand side of the inequality to the left and complete the square. It follows that for suitable constants  $c$ ,

$$(2.21) \quad \begin{aligned} \mathbb{E}\Lambda + h^2 \mathbb{E}\Pi &\leq (\varepsilon + c h^2)^2 + c h^2 (nh)^{-1/2} \log nh \\ &\leq c (h^4 + (nh)^{-1}) \log nh , \end{aligned}$$

and part (b) of the Main Theorem is proved.

In the remainder of the paper we prove the various intermediate results alluded to in the above.

### 3. STRONG CONVEXITY

Here, we prove the inequality (2.8). The first thing to note is that the negative log-likelihood  $L_n(p)$  and the roughness penalization  $\mathbb{R}P(p)$  are both convex functions of  $p$ . It is useful to express the convexity as follows.

First, for nice functions  $p$  and  $\pi$ , the Gateaux variation of  $L_n(p)$  at  $p$  towards  $\pi$  is given by

$$(3.1) \quad \delta L_n(p; \pi - p) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_{in}}{p(x_{in})} - \frac{1 - y_{in}}{1 - p(x_{in})} \right\} \{ \pi(x_{in}) - p(x_{in}) \} ,$$

and then, for all  $p$  and  $\pi$ ,

$$(3.2) \quad L_n(\pi) - L_n(p) - \delta L_n(p; \pi - p) \geq 0 .$$

Thus,  $L_n(p)$  is a convex function of  $p$ . However, it is not strongly convex, i.e., the right hand side of the inequality cannot be replaced by a useful positive (nonnegative) expression.

Secondly, consider the roughness penalization  $\text{RP}(p)$ . Let

$$(3.3) \quad G(p) = \int_0^1 \frac{|p'(t)|^2}{p(t)} dt$$

be the original roughness penalization of GOOD (1971). Consider the function  $f(x, y) = x^2/y$  for  $x \in \mathbb{R}$ ,  $y > 0$ . Its linearization  $\ell(x, y)$  about a point  $(a, b)$  is given by

$$(3.4) \quad \ell(x, y) = \frac{a^2}{b} + \frac{2a}{b}(x - a) - \frac{a^2}{b^2}(y - b)$$

and one verifies that

$$(3.5) \quad f(x, y) - \ell(x, y) = y \left| \frac{x}{y} - \frac{a}{b} \right|^2 .$$

This shows that  $f(x, y)$  is convex in  $x$  and  $y$  jointly.

Observing that

$$\frac{d}{dt} \log \frac{\pi(t)}{p(t)} = \frac{\pi'(t)}{\pi(t)} - \frac{p'(t)}{p(t)} ,$$

we conclude that for nice functions  $\pi$  and  $p$ ,

$$(3.6) \quad G(\pi) - G(p) - \delta G(p; \pi - p) = \int_0^1 \pi(t) \left| \left\{ \log \frac{\pi(t)}{p(t)} \right\}' \right|^2 dt ,$$

where the Gateaux variation is given by

$$(3.7) \quad \delta G(p; \pi - p) = \int_0^1 \frac{2p'(t)}{p(t)} \{ \pi'(t) - p'(t) \} dt - \int_0^1 \left\{ \frac{p'(t)}{p(t)} \right\}^2 \{ \pi(t) - p(t) \} dt .$$

Note that the right hand side of (3.6) in fact equals  $\Delta(\pi, p)$ , see (2.6).

The above may be repeated for the functional  $G(1 - p)$ , showing that

$$(3.8) \quad \mathbf{RP}(\pi) - \mathbf{RP}(p) - \delta\mathbf{RP}(p; \pi - p) = \Pi(\pi, p) ,$$

where  $\Pi(\pi, p)$  is given by (2.7). Since the right hand side of (3.8) is nonnegative, we may conclude that  $\mathbf{RP}(p)$  is convex. Moreover, (3.8) expresses a useful form of strong convexity.

Proceeding along our merry way, let

$$(3.9) \quad M(p) = L_n(p) + h^2 \mathbf{RP}(p) .$$

Setting  $\pi = p_h$  and  $p = p^{nh}$  in (3.2) and (3.8), we obtain

$$(3.10) \quad M(p_h) - M(p^{nh}) - \delta M(p^{nh}; p_h - p^{nh}) \geq h^2 \Pi(p^{nh}, p_h) .$$

However, since  $p^{nh}$  solves (1.8), then  $\delta M(p^{nh}; \pi - p) \geq 0$  for all nice functions  $\pi$  satisfying the constraints, in particular for  $\pi = p_h$ . Thus, (3.10) implies that

$$(3.11) \quad M(p_h) - M(p^{nh}) \geq h^2 \Pi(p^{nh}, p_h) .$$

Finally,

$$M(p_h) - M(p^{nh}) = L_n(p_h) - L_n(p^{nh}) + h^2 \{ \mathbf{RP}(p_h) - \mathbf{RP}(p^{nh}) \} ,$$

and it is easy to verify that

$$L_n(p_h) - L_n(p^{nh}) = S_n(p_h, p^{nh}) - \Lambda_n(p^{nh}, p_h) ,$$

with  $S_n$  defined in (2.9). All of this shows that

$$\Lambda_n(p_h, p^{nh}) + h^2 \Pi(p_h, p^{nh}) \leq S_n(p_h, p^{nh}) + h^2 \{ \mathbf{RP}(p_h) - \mathbf{RP}(p^{nh}) \} .$$

## 4. REPRODUCING KERNEL HILBERT SPACES

We now consider the expectation of  $S_n(p_h, p^{nh})$ . The proper setting for this is that of reproducing kernel Hilbert spaces. In this section, we explore the necessary set-up.

The relevant function space is

$$(4.1) \quad \mathfrak{H} = \left\{ f \in C[0, 1] \left| \begin{array}{l} f \text{ absolutely continuous,} \\ f' \text{ square integrable} \end{array} \right. \right\} ,$$

with the inner products

$$(4.2) \quad \langle f, g \rangle_h = \int_0^1 f(t)g(t) dt + h^2 \int_0^1 f'(t)g'(t) dt ,$$

and norms  $\| \cdot \|_h$  defined via  $\| f \|_h^2 = \langle f, f \rangle_h$  or

$$(4.3) \quad \| f \|_h^2 = \| f \|^2 + h^2 \| f' \|^2 .$$

Here,  $\| \cdot \|$  is the usual  $L^2(0, 1)$  norm.

It is well-known that  $\mathfrak{H}$  is a reproducing kernel Hilbert space for each of the inner products (4.2), i.e., for all  $0 \leq t \leq 1$ , there exist  $\mathfrak{R}_h(t, \cdot) \in \mathfrak{H}$  such that for all  $f \in \mathfrak{H}$ ,

$$(4.4) \quad f(t) = \langle f, \mathfrak{R}_h(t, \cdot) \rangle_h ,$$

and

$$(4.5) \quad \| \mathfrak{R}_h(t, \cdot) \|_h = \mathcal{O}(h^{-1/2}) ,$$

see, e.g., EUBANK (1999). This is all that is needed for the rkhs trick.

**The reproducing kernel Hilbert space trick.** In its simplest form, the reproducing kernel Hilbert space trick we have in mind deals with sums of the form

$$(4.6) \quad \mathcal{S}_n(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} f(x_{in}) ,$$

where  $\varepsilon_{1,n}, \varepsilon_{2,n}, \dots, \varepsilon_{nn}$  are uncorrelated random variables with

$$\mathbb{E}[\varepsilon_{in}] = 0 \quad , \quad \mathbb{E}[|\varepsilon_{in}|^2] \leq 1 ,$$

and  $f \in \mathfrak{H}$  is a *random function*. We now prove

(4.7) LEMMA. *There exists a constant  $c$  such that for all random or deterministic  $f \in \mathfrak{H}$  and  $0 < h \leq 1$ ,*

$$\mathbb{E}[|\mathcal{S}_n(f)|] \leq c(nh)^{-1/2} \{ \mathbb{E}[\|f\|_h^2] \}^{1/2} .$$

PROOF. Write  $f(x_{in}) = \langle \mathfrak{R}_h(x_{in}, \cdot), f \rangle_h$  and obtain

$$\begin{aligned} \mathcal{S}_n(f) &= \left\langle \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} \mathfrak{R}_h(x_{in}, \cdot), f \right\rangle_h \\ (4.8) \qquad &\leq \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} \mathfrak{R}_h(x_{in}, \cdot) \right\|_h \|f\|_h . \end{aligned}$$

Since

$$\mathbb{E} \left[ \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} \mathfrak{R}_h(x_{in}, \cdot) \right\|_h^2 \right] = \left( \frac{1}{n} \right)^2 \sum_{i=1}^n \mathbb{E}[|\varepsilon_{in}|^2] \left\| \mathfrak{R}_h(x_{in}, \cdot) \right\|_h^2 ,$$

the lemma follows. Q.e.d.

## 5. THE REPRODUCING KERNEL HILBERT SPACE TRICK FOR LOGISTIC REGRESSION

After the preliminary work of the previous section, we are now ready to prove the following bound on  $S_n(p_h, p^{nh})$ , defined in (2.9),

$$(5.1) \quad \left\{ \mathbb{E}[|S_n(p_h, p^{nh})|] \right\}^2 \leq c h^4 (nh)^{-1} (\log nh)^3 + c(nh)^{-1} \left\{ \mathcal{Q}\Lambda(p_h, p^{nh}) + h^2 \Pi(p_h, p^{nh}) \right\} ,$$

with  $\Pi(p, \pi)$  given by (2.7) and  $\mathcal{Q}\Lambda(p, \pi)$  by (2.3).

We must apply the reproducing kernel Hilbert space trick to the sums  $S_n(p_h, p^{nh})$ . For ease of exposition, write  $S_n(p_h, p^{nh})$  as

$$(5.2) \quad S_n(p_h, p^{nh}) = \mathbb{T}_n + \mathbb{U}_n ,$$

where

$$(5.3) \quad \begin{aligned} \mathbb{T}_n &= \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} \log \frac{p_h(x_{in})}{p^{nh}(x_{in})} , \\ \mathbb{U}_n &= \frac{1}{n} \sum_{i=1}^n \varepsilon_{in} \log \frac{1 - p_h(x_{in})}{1 - p^{nh}(x_{in})} , \end{aligned}$$

with  $\varepsilon_{in} = y_{in} - p_h(x_{in})$ .

It suffices to prove the following bounds on  $\mathbb{T}_n$  and  $\mathbb{U}_n$ ,

(5.4) LEMMA. *Suppose  $p_o$  satisfies (1.15)-(1.16). There exists a constant  $c$  such that for all  $n$  and  $0 < h < 1$  with  $nh \rightarrow \infty$ ,*

$$\begin{aligned} \{ \mathbb{E}[|\mathbb{T}_n(p_h, p^{nh})|] \}^2 &\leq c h^4 (nh)^{-1} (\log nh)^2 + \\ &\quad c (nh)^{-1} \{ \text{QKL}(p_h, p^{nh}) + h^2 \Delta(p_h, p^{nh}) \} , \\ \{ \mathbb{E}[|\mathbb{U}_n(p_h, p^{nh})|] \}^2 &\leq c h^4 (nh)^{-1} (\log nh)^2 + \\ &\quad c (nh)^{-1} \{ \text{QKL}(1 - p_h, 1 - p^{nh}) + h^2 \Delta(1 - p_h, 1 - p^{nh}) \} . \end{aligned}$$

We only prove the result for  $\mathbb{T}_n$ . The treatment of  $\mathbb{S}_n$  is similar.

The first step is to replace the  $\varepsilon_{in}$  with

$$(5.5) \quad \delta_{in} = y_{in} - p_o(x_{in}) , \quad i = 1, 2, \dots, n ,$$

so define

$$(5.6) \quad T_n = \frac{1}{n} \sum_{i=1}^n \delta_{in} \log \frac{p_h(x_{in})}{p^{nh}(x_{in})} .$$

The bound for  $\mathbb{T}_n$  now follows from Lemmas (5.7) and (5.8).

(5.7) LEMMA. *For all  $n$  and  $h$ , with  $nh > 2$ ,*

$$|\mathbb{T}_n - T_n|^2 \leq (nh)^{-1} \text{QKL}_n(p_h, p^{nh}) .$$

PROOF OF LEMMA (5.7). For the difference we have

$$|\mathbb{T}_n - T_n| \leq \max \frac{|p_o(x_{in}) - p_h(x_{in})|}{\sqrt{p_h(x_{in})}} \cdot \frac{1}{n} \sum_{i=1}^n \sqrt{p_h(x_{in})} \left| \log \frac{p_h(x_{in})}{p_o(x_{in})} \right| .$$

By the truncation (1.9), the maximum term is bounded by  $(nh)^{-1/2}$  (the numerator is at *most*  $(nh)^{-1}$  and the denominator is at *least*  $(nh)^{-1/2}$ ). With Cauchy-Schwarz, the sum is bounded by  $\{ \text{QKL}_n(p_h, p^{nh}) \}^{1/2}$ . Q.e.d.

We now study  $T_n$  using reproducing kernel Hilbert space tricks.

(5.8) LEMMA. *Suppose that  $p_o$  satisfies (1.15)-(1.16). There exists a constant such that for  $nh \rightarrow \infty$ ,*

$$\begin{aligned} \{ \mathbb{E}[T_n] \}^2 &\leq c h^4 (nh)^{-1} (\log nh)^2 + \\ &\quad c (nh)^{-1} \{ \mathbb{E}[\text{QKL}(p_h, p^{nh})] + h^2 \mathbb{E}[\Delta(p_h, p^{nh})] \} . \end{aligned}$$

PROOF. First, note that  $p^{nh}$  and  $p_h$  both belong to  $\mathfrak{H}$  and since they are bounded away from zero and one, then  $\sqrt{p_h} \log(p_h/p^{nh}) \in \mathfrak{H}$  as well. Now, rewrite  $T_n$  as

$$(5.9) \quad T_n = \frac{1}{n} \sum_{i=1}^n \frac{\delta_{in}}{\sqrt{p_h(x_{in})}} \times \sqrt{p_h(x_{in})} \log \frac{p^{nh}(x_{in})}{p_h(x_{in})}$$

and apply (4.8) to obtain

$$(5.10) \quad \begin{aligned} T_n &= \left\langle \frac{1}{n} \sum_{i=1}^n \frac{\delta_{in} \mathfrak{R}_h(x_{in}, \cdot)}{\sqrt{p_h(x_{in})}}, \sqrt{p_h} \log \frac{p_h}{p^{nh}} \right\rangle_h \\ &\leq \left\| \frac{1}{n} \sum_{i=1}^n \frac{\delta_{in} \mathfrak{R}_h(x_{in}, \cdot)}{\sqrt{p_h(x_{in})}} \right\|_h \left\| \sqrt{p_h} \log \frac{p_h}{p^{nh}} \right\|_h . \end{aligned}$$

Now, the expectation of the square of the first factor equals

$$(5.11) \quad \left( \frac{1}{n} \right)^2 \sum_{i=1}^n \frac{p_o(x_{in})(1 - p_o(x_{in}))}{p_h(x_{in})} \left\| \mathfrak{R}_h(x_{in}, \cdot) \right\|_h^2 .$$

By the shrinking (1.9), the fraction in the sum is well behaved. The squared norm on the right is  $\mathcal{O}(h^{-1})$ , see (4.5). Thus, the expression in (5.11) is  $\mathcal{O}((nh)^{-1})$ .

The square of the second factor in (5.10) equals

$$(5.12) \quad \left\| \sqrt{p_h} \log \frac{p_h}{p^{nh}} \right\|^2 + h^2 \left\| \left\{ \sqrt{p_h} \log \frac{p_h}{p^{nh}} \right\}' \right\|^2 .$$

The first term is just  $\text{QKL}(p_h, p^{nh})$ , and the second term is bounded by

$$(5.13) \quad 2 h^2 \left\| \left\{ \sqrt{p_h} \right\}' \log \frac{p_h}{p^{nh}} \right\|^2 + 2 h^2 \left\| \sqrt{p_h} \left\{ \log \frac{p_h}{p^{nh}} \right\}' \right\|^2 .$$

The second term of *this* equals  $2h^2 \Delta(p_h, p^{nh})$ , see (2.6). Explicitly writing out the first term, we have by Cauchy-Schwarz,

$$2h^2 \int_0^1 \frac{|p'_h(t)|^2}{p_h(t)} \left| \log \frac{p_h(t)}{p^{nh}(t)} \right|^2 dt \leqslant 2h^2 \left| \int_0^1 \frac{|p'_h(t)|^4}{|p_h(t)|^3} dt \right|^{1/2} \left| \int_0^1 p_h(t) \left| \log \frac{p_h(t)}{p^{nh}(t)} \right|^4 dt \right|^{1/2} .$$

Now, by the truncation (1.9), the first factor essentially does not exceed

$$\left| \int_0^1 \frac{|p'_o(t)|^4}{|p_o(t)|^3} dt \right|^{1/2} ,$$

and this is finite by virtue of the conditions (1.15)-(1.16), cf. EGGERMONT and LARICCIA (2001), Exercise (4.6.4).

By the constraints (1.8) and (1.9), the second factor is bounded by

$$\log nh \{ \text{QKL}(p_h, p^{nh}) \}^{1/2} ,$$

and then for appropriate constants  $c$ , the first term of (5.13) may be bounded by

$$ch^2 \log nh \{ \text{QKL}(p_h, p^{nh}) \}^{1/2} \leqslant 2ch^4 (\log nh)^2 + 2c \text{QKL}(p_h, p^{nh}) .$$

Putting all this together in (5.10), we get the bound for appropriate constants  $c$ ,

$$\{ \mathbb{E}[T_n] \}^2 \leqslant c(nh)^{-1} \{ \mathbb{E}[ \text{QKL}(p_h, p^{nh}) ] + h^2 \mathbb{E}[ \Delta(p_h, p^{nh}) ] \} + ch^4 (nh)^{-1} (\log nh)^2 ,$$

thus proving the lemma.

Q.e.d.

## 6. AN ELEMENTARY INEQUALITY

Here, we prove the inequality connecting the “quadratic” and the plain Kullback-Leibler distance for the “truncated” functions.

(6.1) AN ELEMENTARY INEQUALITY. *Let  $b > 1$ . For  $0 \leq r \leq b$ ,*

$$\frac{r |\log r|^2}{r \log r + 1 - r} \leq C(b) \log b ,$$

with  $C(b) \rightarrow 1$  as  $b \rightarrow \infty$ .

The proof is delayed until the end of the section.

(6.2) LEMMA. *Let  $b > 1$ . If  $p$  and  $\pi$  are nice functions satisfying*

$$p(t), \pi(t) \in [b^{-1}, 1] \quad \text{for all } t \in [0, 1] ,$$

then, there exists a constant  $c$  such that for  $b \rightarrow \infty$ ,

$$\mathbf{QKL}(p, \pi) \leq c \mathbf{KL}(p, \pi) \log b , \quad \mathbf{QKL}_n(p, \pi) \leq c \mathbf{KL}_n(p, \pi) \log b ,$$

PROOF OF LEMMA (6.2). Let  $x = p(t)$ ,  $y = \pi(t)$  and  $r = x/y$ . Then,  $0 \leq r \leq b$  and

$$x \left| \log \frac{x}{y} \right|^2 = y r \left| \log r \right|^2 \leq C(b) y (r \log r + 1 - r) \log b ,$$

by Lemma (6.1). Since  $r = x/y$ , multiplication by  $y$  yields

$$x \left| \log \frac{x}{y} \right|^2 \leq C(b) (x \log \frac{x}{y} + y - x) \log b .$$

With the identification  $x = p(t)$  and  $y = \pi(t)$ , the lemma follows upon integration. With  $t = x_{in}$  and summing over  $i$ , the discrete analogue is provided for. Q.e.d.

PROOF OF LEMMA (6.1). For  $r > 0$ ,  $r \neq 1$ , define

$$\varphi(r) = \frac{r |\log r|^2}{r \log r + 1 - r} ,$$

and set  $\varphi(1) = 1$ . Then  $\varphi(r)$  is continuous for  $r > 0$ , and, for  $r \neq 1$ , continuously differentiable. It now suffices to show that  $\log \varphi(r)$  is an increasing function.

One verifies that

$$\frac{d}{dr} \log \varphi(r) = \frac{(r+1) \log r + 2 - 2r}{r \log r (r \log r + 1 - r)}, \quad r \neq 1.$$

After dividing numerator and denominator by  $\log r$ , we see that the derivative is positive for all  $r$  if and only if

$$r + 1 - \frac{2(r-1)}{\log r} > 0 \quad \text{for all } r \neq 1.$$

This is equivalent to

$$(6.3) \quad \frac{1}{1+r} - \frac{\log r}{2(r-1)} < 0 \quad \text{for all } r \neq 1.$$

For  $r > 1$ , the above inequality is equivalent to

$$(6.4) \quad \frac{r-1}{r+1} - \frac{1}{2} \log r < 0 \quad \text{for all } r > 1.$$

Since the expression on the left vanishes for  $r = 1$ , and its derivative

$$\frac{-(r-1)^2}{2r(r+1)^2},$$

is negative, then (6.4) follows.

For  $r < 1$ , the statement (6.3) is equivalent to

$$(6.5) \quad \frac{r-1}{r+1} - \frac{1}{2} \log r > 0 \quad \text{for all } r < 1,$$

and the same argument as above shows that this holds also.

Thus, (6.3) holds for all  $r \neq 1$ .

Q.e.d.

## 7. SUMS VS INTEGRALS

A technical and annoying detail is the occurrence of the discrete sums and integrals like  $\text{KL}_n(p_h, p^{nh})$  and  $\text{KL}(p_h, p^{nh})$ .

(7.1) QUADRATURE LEMMA. *Let  $x_{in}, i = 1, 2, \dots, n$ , be a asymptotically uniform design. If  $p$  and  $\pi$  are nice functions, then*

$$\left| \Lambda_n(p, \pi) - \Lambda(p, \pi) \right|^2 \leq c n^{-2} \{ \text{RP}(p) \text{QKL}(p, \pi) + \Pi(p, \pi) \} .$$

PROOF. It suffices to prove the result for

$$S_n = \frac{1}{n} \sum_{i=1}^n p(x_{in}) \log \frac{p(x_{in})}{\pi(x_{in})} \quad \text{and} \quad S = \int_0^1 p(t) \log \frac{p(t)}{\pi(t)} .$$

By the asymptotic uniformity of the design,

$$| S_n - S | \leq c n^{-1} \left\| \left\{ p \log \frac{p}{\pi} \right\}' \right\|_1 .$$

Using the triangle inequality, the last norm is bounded by

$$(7.2) \quad \left\| p' \log \frac{p}{\pi} \right\|_1 + \left\| p \left\{ \log \frac{p}{\pi} \right\}' \right\|_1 .$$

The Cauchy-Schwarz inequality gives the bounds

$$\begin{aligned} \left\| p' \log \frac{p}{\pi} \right\|_1^2 &\leq \int_0^1 \frac{|p'(t)|^2}{p(t)} dt \cdot \int_0^1 p(t) \left| \log \frac{p(t)}{\pi(t)} \right|^2 dt \\ &= G(p) \text{QKL}(p, \pi) , \end{aligned}$$

and

$$\left\| p \left\{ \log \frac{p}{\pi} \right\}' \right\|_1^2 \leq \| p \|_1 \Delta(p, \pi) \leq \Delta(p, \pi),$$

since  $p \leq 1$  everywhere. The lemma follows. Q.e.d.

## 8. THE BIAS DUE TO THE ROUGHNESS PENALIZATION

Finally, we consider the term  $h^2 \{ \text{RP}(p_h) - \text{RP}(p^{nh}) \}$  in the strong convexity inequality (2.12).

PROOF OF (2.18). It is helpful to write  $\text{RP}(p)$  as

$$(8.1) \quad \text{RP}(p) = G(p) + G(1-p) ,$$

with  $G(p)$  as is (3.3).

First, to simplify notation, set  $q \equiv p^{nh}$ . Then, by convexity,

$$(8.2) \quad \begin{aligned} G(p_h) - G(q) &= \left\| \{ \sqrt{p_h} \}' \right\|^2 - \left\| \{ \sqrt{q} \}' \right\|^2 \\ &\leq 2 \left\langle \{ \sqrt{p_h} \}', \{ \sqrt{p_h} \}' - \{ \sqrt{q} \}' \right\rangle, \end{aligned}$$

where  $\langle \cdot, \cdot \rangle$  denotes the usual  $L^2(0, 1)$  inner product. Now, since  $p_o$  satisfies (1.14)-(1.16), so does  $p_h$ , and integration by parts gives

$$(8.3) \quad \begin{aligned} \left\langle \{ \sqrt{p_h} \}', \{ \sqrt{p_h} \}' - \{ \sqrt{q} \}' \right\rangle &= - \left\langle \{ \sqrt{p_h} \}'', \sqrt{p_h} - \sqrt{q} \right\rangle \\ &\leq \left\| \{ \sqrt{p_h} \}'' \right\| \left\| \sqrt{p_h} - \sqrt{q} \right\|. \end{aligned}$$

Now, recall that  $p_h$  is the shrunken version of  $p_o$ , so that

$$\left\| \{ \sqrt{p_h} \}'' \right\| \approx \left\| \{ \sqrt{p_o} \}'' \right\| < \infty, \quad \text{for } nh \rightarrow \infty.$$

Finally, the inequality  $\left\| \sqrt{p_h} - \sqrt{q} \right\|^2 \leq \text{KL}(p_h, q)$  clinches the argument. Q.e.d.

## 9. BOUNDED LOG ODDS RATIOS

We now show that the factors  $\log n$  in Theorem (1.17) are not necessary in case of bounded log odds ratios, i.e.,

$$(9.1) \quad \left| \log \frac{p_o(t)}{1 - p_o(t)} \right| \leq K, \quad t \in [0, 1],$$

for some constant  $K$ . Put differently, we are assuming that there exists a positive constant  $\delta$  such that

$$(9.2) \quad \delta \leq p_o(t) \leq 1 - \delta, \quad t \in [0, 1].$$

We state the precise results.

(9.3) THEOREM. *If  $p_o$  satisfies (1.14) and (9.2), then the solution  $\pi^{nh}$  of (1.4) satisfies*

$$\mathbb{E} \left[ \Lambda(p_o, \pi^{nh}) \right] = \mathcal{O}(n^{-2/3}),$$

provided  $h \asymp n^{-1/3}$  (deterministically).

If, moreover,  $p_o$  satisfies (1.15)-(1.16), then the rate improves to

$$\mathbb{E} \left[ \Lambda(p_o, \pi^{nh}) \right] = \mathcal{O}(n^{-4/5}),$$

provided  $h \asymp n^{-1/5}$ , again deterministically.

Oddly enough, the result at the basis of it all is the following almost sure version of Theorem (1.17).

(9.4) THEOREM. *If  $p_o$  satisfies (1.14) and (9.2), then the solution  $p^{nh}$  of (1.8) satisfies*

$$\Lambda(p_h, p^{nh}) + h^2 \Pi(\pi_h, p^{nh}) = \mathcal{O}(n^{-2/3}(\log n)^2), \quad \text{almost surely.}$$

provided  $h \asymp n^{-1/3}$  (deterministically).

If, moreover,  $p_o$  satisfies (1.15)-(1.16), then the rate improves to

$$\Lambda(p_h, p^{nh}) + h^2 \Pi(\pi_h, p^{nh}) = \mathcal{O}(n^{-4/5}(\log n)^2), \quad \text{almost surely.}$$

provided  $h \asymp n^{-1/5}$ , again deterministically.

PROOF OF THEOREM(9.3). We only consider the “4/5” case. Let

$$(9.5) \quad \ell^{nh}(t) = \log \frac{p_h(t)}{p^{nh}(t)}, \quad t \in [0, 1].$$

Theorem (9.4) implies that

$$(9.6) \quad \|\ell^{nh}\|_h^2 = \mathcal{O}(n^{-4/5}(\log n)^2) \quad \text{almost surely,}$$

for  $h \asymp n^{-1/5}$ . In combination with (4.4)-(4.5), we then get

$$(9.7) \quad \|\ell^{n,h}\|_\infty^2 = \mathcal{O}(n^{-3/5}(\log n)^2) \quad \text{almost surely,}$$

(we lost a factor  $h^{-1}$ ), so that

$$(9.8) \quad \|p^{nh} - p_h\|_\infty \longrightarrow 0 \quad \text{almost surely.}$$

Now, consider the problem

$$(9.9) \quad \begin{aligned} & \text{minimize} && L_n(p) + h^2 \text{RP}(p) \\ & \text{subject to} && \text{RP}(p) < \infty, \\ & && \frac{1}{2} \delta \leq p \leq 1 - \frac{1}{2} \delta. \end{aligned}$$

Denote the solution by  $p = \varpi^{nh}$ . Redoing the proof of Theorem (1.17) shows that the factors  $\log(nh)$  are replaced by  $\log(\delta/2)$ , which are just constants. Thus,

$$(9.10) \quad \mathbb{E} \left[ \Lambda(p_h, \varpi^{nh}) \right] = \mathcal{O}(n^{-4/5}),$$

for  $h \asymp n^{-1/5}$ . By Theorem (9.4), similar to the reasoning that lead to (9.8),  $\varpi^{nh}$  and  $p^{nh}$  coincide almost surely as  $n \rightarrow \infty$ . Thus, the constraints in (9.9) are not active, and  $\varpi^{nh}$  coincides almost surely with  $\pi^{nh}$ , the solution of (1.4).

Finally, since  $p_o$  is bounded away from zero and one, it is an easy exercise to show that for all  $p$  satisfying the constraint (9.2),

$$\Lambda(p_h, p) - \Lambda(p_o, p) = \mathcal{O}((nh)^{-1}),$$

so that (9.10) implies the bound of the theorem. Q.e.d.

PROOF OF THEOREM (9.4). Inspection of the proof of Theorem (1.17) reveals that it suffices to get almost sure bounds in §5. In particular, it suffices to show that, almost surely,

$$(9.11) \quad \left\| \frac{1}{n} \sum_{i=1}^n \frac{\delta_{in} \mathfrak{R}_h(x_{in}, \cdot)}{\sqrt{p_h(x_{in})}} \right\|_h = \mathcal{O} \left( \left\{ \frac{\log n}{nh} \right\}^{1/2} \right),$$

where  $\delta_{in} = y_{in} - p_o(x_{in})$ . One way to prove this is by way of the (martingale) method of bounded differences of MCDIARMID (1989) as implemented by DEVROYE (1996). For (more) readily accessible renditions, see, e.g., DEVROYE, GYÖRFI and LUGOSI (1996) or EGGERMONT and LARICCIA (2001).

Let  $\delta_n = (\delta_{1,n}, \delta_{2,n}, \dots, \delta_{n,n})$ , and define

$$\psi(\delta_n) = \left\| \frac{1}{n} \sum_{i=1}^n \frac{\delta_{in} \mathfrak{R}_h(x_{in}, \cdot)}{\sqrt{p_h(x_{in})}} \right\|_h.$$

We now wish to see how much  $\psi(\delta_n)$  can change when we let  $\delta_{in}$  vary while keeping the other components fixed. Of course, the  $\delta_{in}$  themselves are bounded:  $-1 \leq \delta_{in} \leq 1$ ,  $i = 1, 2, \dots, n$ . To simplify(?) notation somewhat, let  $s \in \mathbb{R}^n$ , and for  $1 \leq i \leq n$  and  $r \in \mathbb{R}$ , define  $s_i(r) \in \mathbb{R}^n$  by

$$(s_i(r))_j = \begin{cases} \delta_{jn}, & j \neq i, \\ r, & j = i. \end{cases}$$

Then, by the triangle inequality, for all  $q, r$ ,

$$\left| \psi(s_i(q)) - \psi(s_i(r)) \right| \leq \frac{1}{n} \left\| \frac{q-r}{\sqrt{p_h(x_{in})}} \mathfrak{K}_h(x_{in}, \cdot) \right\|_h \leq c \frac{|q-r|}{n\sqrt{h}},$$

for an appropriate constant  $c$ , since  $p_o$ , and hence  $p_h$ , is bounded away from 0 and by the bound (4.4) on the reproducing kernel.

It follows that, for the same constant  $c$ , for all  $n$  and  $i = 1, 2, \dots, n$ ,

$$(9.12) \quad \sup_{0 \leq q, r \leq 1} \left| \psi(s_i(q)) - \psi(s_i(r)) \right| \leq c n^{-1} h^{-1/2}.$$

Now, McDiarmid's theorem implies that for all  $t > 0$ ,

$$\mathbb{P}[|\psi(\delta_n) - \mathbb{E}[\psi(\delta_n)]| > t] \leq 2 \exp(-\frac{1}{2} n h t^2 / c^2),$$

with the same constant  $c$  as in (9.12). With the Borel-Cantelli theorem, then

$$\psi(\delta_n) = \mathbb{E}[\psi(\delta_n)] + \mathcal{O}((nh)^{-1/2} \log n) \quad \text{almost surely.}$$

Since  $\mathbb{E}[\psi(\delta_n)] = \mathcal{O}((nh)^{-1/2})$ , this implies (9.11) and we are done. Q.e.d.

## REFERENCES

- Cox, D.D., O'Sullivan, F. (1990), *Asymptotic analysis of penalized likelihood and related estimators*, Ann. Statist. **18**, 1676–1695.
- Eggermont, P.P.B., LaRiccia, V.N. (1999), *Optimal convergence rates for Good's nonparametric maximum likelihood density estimator*, Ann. Statist. **28**, 1600–1615.

- Eggermont, P.P.B., LaRiccia, V.N. (2001), *Maximum penalized likelihood estimation. Volume 1: Density estimation*, Springer-Verlag, New York.
- Devroye, L. (1996), *Exponential inequalities in nonparametric estimation*, Nonparametric functional estimation and related topics (G. Roussas, ed.), Kluwer, Dordrecht, pp. 31–44.
- Devroye, L., Györfi, L., Lugosi, G. (1996), *A probabilistic theory of pattern recognition*, Springer-Verlag, New York.
- Eubank, R.L. (1999), *Spline smoothing and nonparametric regression. Second edition*, Marcel Dekker, New York.
- Finney, D.J. (1992), *Statistical Method in Bioassay. Third Edition*, Oxford University Press, Oxford.
- Good, I.J. (1971), *A nonparametric roughness penalty for probability densities*, Nature **229**, 29–30.
- Gu, C., Qiu, C. (1993), *Smoothing spline density estimation: Theory*, Ann. Statist. **21**, 217–234.
- Gu, C. (2002), *Smoothing spline ANOVA models*, Springer-Verlag, New York.
- Kass, R.E. (1989), *The geometry of asymptotic inference*, Statistical Science **4**, 188–219.
- Kleinbaum, D.G., Klein, M. (2002), *Logistic regression. Second Edition*, Springer-Verlag, New York.
- Kemperman, J.H.B. (1967), *On the optimum rate of transmitting information*, Springer Lecture Notes in Mathematics **23**, 126–169.
- Kooperberg, C., Stone, C.J. (1991), *A study of log-spline density estimation*, Comput. Statist. Data Anal. **12**, 327–347.
- McDiarmid, C. (1989), *On the method of bounded differences*, Surveys in combinatorics 1989, Cambridge University Press, Cambridge, pp. 148–188.
- Stone, C.J. (1982), *Optimal global rates of convergence for nonparameteric regression*, Ann. Statist. **10**, 1040–1053.

FOOD AND RESOURCE ECONOMICS  
 213 TOWNSEND HALL · 531 SOUTH COLLEGE AVENUE  
 UNIVERSITY OF DELAWARE · NEWARK, DELAWARE 19717-1303