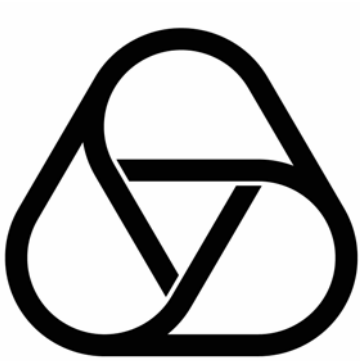
Oberwolfach Preprints



OWP 2012 - 07

LÁSZLÓ GYÖRFI, HARRO WALK

# Strongly Consistent Density Estimation of Regression Residual

Mathematisches Forschungsinstitut Oberwolfach gGmbH
Oberwolfach Preprints (OWP) ISSN 1864-7596

## Oberwolfach Preprints (OWP)

Starting in 2007, the MFO publishes a preprint series which mainly contains research results related to a longer stay in Oberwolfach. In particular, this concerns the Research in Pairs-Programme (RiP) and the Oberwolfach-Leibniz-Fellows (OWLF), but this can also include an Oberwolfach Lecture, for example.

A preprint can have a size from 1 - 200 pages, and the MFO will publish it on its website as well as by hard copy. Every RiP group or Oberwolfach-Leibniz-Fellow may receive on request 30 free hard copies (DIN A4, black and white copy) by surface mail.

Of course, the full copy right is left to the authors. The MFO only needs the right to publish it on its website *www.mfo.de* as a documentation of the research work done at the MFO, which you are accepting by sending us your file.

In case of interest, please send a **pdf file** of your preprint by email to *rip@mfo.de* or *owlf@mfo.de*, respectively. The file should be sent to the MFO within 12 months after your stay as RiP or OWLF at the MFO.

There are no requirements for the format of the preprint, except that the introduction should contain a short appreciation and that the paper size (respectively format) should be DIN A4, "letter" or "article".

On the front page of the hard copies, which contains the logo of the MFO, title and authors, we shall add a running number (20XX - XX).

We cordially invite the researchers within the RiP or OWLF programme to make use of this offer and would like to thank you in advance for your cooperation.

## Imprint:

Mathematisches Forschungsinstitut Oberwolfach gGmbH (MFO)
Schwarzwaldstrasse 9-11
77709 Oberwolfach-Walke
Germany

Tel +49 7834 979 50
Fax +49 7834 979 55
Email admin@mfo.de
URL www.mfo.de

The Oberwolfach Preprints (OWP, ISSN 1864-7596) are published by the MFO.

# Strongly consistent density estimation of regression residual[☆]

László Györfi[a,*], Harro Walk[b]

[a] *Department of Computer Science and Information Theory, Budapest University of Technology and Economics, 1521 Stoczek u. 2, Budapest, Hungary*
[b] *Department of Mathematics, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany*

**Abstract**

Consider the regression problem with a response variable $Y$ and with a $d$-dimensional feature vector $X$. For the regression function $m(x) = \mathbb{E}\{Y|X = x\}$, this paper investigates methods for estimating the density of the residual $Y - m(X)$ from independent and identically distributed data. For heteroscedastic regression, we prove the strong universal (density-free) $L_1$-consistency of a recursive and a nonrecursive kernel density estimate based on a regression estimate.

*Keywords:* Regression residual, nonparametric kernel density estimation, nonparametric regression estimation, heteroscedastic regression.
*2000 MSC:* primary 62G07, secondary 62G20

## 1. Introduction

Let $Y$ be a real valued random variable and let $X = (X^{(1)}, \dots, X^{(d)})$ be a $d$-dimensional random vector. The coordinates of $X$ may have different types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. In the sequel we do not assume anything about the distribution of $X$. The task of regression analysis is to estimate $Y$ given $X$, i.e., one aims to find a function $F$ defined on the range of $X$ such that $F(X)$ is "close" to $Y$. Typically, closeness is measured in terms of the *mean squared error* of $F$,

$$\mathbb{E}\{(F(X) - Y)^2\}.$$

It is well-known that the mean squared error is minimized by the regression function $m$ with

$$m(x) = \mathbb{E}\{Y \mid X = x\}, \tag{1}$$

since, for each measurable function $F$, the mean squared error can be decomposed into

$$\mathbb{E}\{(F(X) - Y)^2\} = \mathbb{E}\{(m(X) - Y)^2\} + \int_{\mathbb{R}^d} (m(x) - F(x))^2 \mu(dx),$$

[☆] This research was supported through the programme "Research in Pairs" by the Mathematisches Forschungsinstitut Oberwolfach in 2012.

[*] Corresponding author
*Email addresses:* gyorfi@cs.bme.hu (László Györfi), walk@mathematik.uni-stuttgart.de (Harro Walk)

where $\mu$ denotes the distribution of $X$. The second term on the right hand side is called *excess error* or integrated squared error of the function $F$. Clearly, the mean squared error of $F$ is close to its minimum if and only if the excess error $\int_{\mathbb{R}^d}(m(x)-F(x))^2\mu(dx)$ is close to zero.

The regression function cannot be calculated as long as the distribution of $(X,Y)$ is unknown. Assume, however, that we observed data

$$D_n = \{(X_1,Y_1),\dots,(X_n,Y_n)\} \tag{2}$$

consisting of independent and identically distributed copies of $(X,Y)$. $D_n$ can be used to produce an estimate $m_n = m_n(\cdot, D_n)$ of the regression function $m$. Since $m$ arises from $L_2$ considerations, it is natural to study $L_2(\mu)$ convergence of the regression estimate $m_n$ to $m$. In particular, the estimator $m_n$ is called *strongly universally consistent* if its excess error satisfies

$$\int_{\mathbb{R}^d}(m(x)-m_n(x))^2\mu(dx) \to 0 \text{ a.s.}$$

for all distributions of $(X,Y)$ with $\mathbb{E}|Y|^2 < \infty$. (Cf. Györfi et al. [13].)

It is of great importance to be able to estimate the various characteristics of the residual

$$Y - m(X).$$

In this paper we deal with the problem how to estimate the density $f$ of the residual $Y-m(X)$, assuming that the density $f$ exists. Our aim is to estimate $f$ from i.i.d. data (2).

Under some smoothness conditions on the density $f$, Ahmad [1], Cheng [4], [3], Efromovich [9], [10], Akritas and Van Keilegom [2], Neumeyer and Van Keilgom [14] studied the estimate the density of the residual. In the model of independent measurement error $Z$ one has

$$Y = m(X) + Z \tag{3}$$

such that $\mathbb{E}\{Z\} = 0$, and $X$ and $Z$ are independent. Sometimes it is called additive noise model or homeoscedastic regression model. Under the additive noise model (3), Devroye et al. [7] introduced a density estimate of the residual, and proved its universal (density free) strong consistency in $L_1$.

## 2. A recursive kernel estimate

In this paper we extend the result of Devroye et al. [7] such that don't assume the additive noise model (3), i.e., consider heteroscedastic regression problem. We only assume that, for given $X = x$, the conditional density of the residual $Y - m(X)$ exists. This conditional density is denoted by $f(z \mid x)$. Then

$$f(z) = \int_{\mathbb{R}^d} f(z \mid x)\mu(dx).$$

Suppose that based on the data $(X_1,Y_1),\dots,(X_n,Y_n)$, we are given a strongly universally consistent regression estimate $m_n$. We introduce a recursive density estimate

of the residual, which is a slight modification of the recursive kernel density estimate proposed by Wolverton and Wagner [17] and Yamato [16] for observable i.i.d. random variables $Y_i - m(X_i)$. Let $K$ be a density on $\mathbb{R}$, called kernel, $\{h_i\}$ is the bandwidth sequence. For a bandwidth $h > 0$, introduce the notation

$$K_h(z) = \frac{1}{h} K(z/h).$$

Then the recursive kernel estimate is defined by

$$f_n(z) := \frac{1}{n} \sum_{i=1}^{n} K_{h_i}(z - Z_i), \tag{4}$$

where in the $i$-th term we plug-in the approximation of the $i$-th residual

$$Z_i := Y_i - m_{i-1}(X_i).$$

Notice that the estimate $f_n$ defined by (4) can be calculated sequentially: Put $f_0 = 0$ and $m_0 = 0$, then for $n \geq 1$, we have that

$$f_n(z) = \left(1 - \frac{1}{n}\right) f_{n-1}(z) + \frac{1}{n} K_{h_n}\left(z - (Y_n - m_{n-1}(X_n))\right).$$

**Theorem 1.** *Assume that $Y$ is square integrable. Suppose that we are given a strongly universally consistent regression estimate $m_n$, i.e.,*

$$\int_{\mathbb{R}^d} (m(x) - m_n(x))^2 \mu(dx) \to 0 \ \textit{a.s.}$$

*and for given $X = x$, the conditional density of the residual $Y - m(X)$ exists. Assume that the kernel function $K$ is a square integrable density, and*

$$\lim_{n\to\infty} h_n = 0 \ \textit{and} \ \sum_{n=1}^{\infty} \frac{1}{n^2 h_n} < \infty. \tag{5}$$

*Then*

$$\lim_{n\to\infty} \int_{\mathbb{R}} |f_n(z) - f(z)| dz = 0$$

*a.s.*

Proof For given $X = x$ and for given $(X_1, Y_1), \ldots, (X_n, Y_n)$, the approximate residual

$$Y - m_n(X) = Y - m(X) + m(X) - m_n(X)$$

has the conditional density $f(z + m_n(x) - m(x) \mid x)$ and so the density $g_n(z)$ of $Y - m_n(X)$ can be calculated as follows:

$$g_n(z) = \int_{\mathbb{R}^d} f(z + m_n(x) - m(x) \mid x) \mu(dx).$$

Next we show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} |g_n(z) - f(z)| dz = 0 \tag{6}$$

a.s. For $\delta > 0$, introduce the notation

$$\Delta_x(\delta) := \sup_{|u| \le \delta} \int_{\mathbb{R}} |f(z+u \mid x) - f(z \mid x)| dz.$$

Thus,

$$\begin{aligned}
& \int_{\mathbb{R}} |g_n(z) - f(z)| dz \\
= & \int_{\mathbb{R}} | \int_{\mathbb{R}^d} f(z + m_n(x) - m(x) \mid x) \mu(dx) - \int_{\mathbb{R}^d} f(z \mid x) \mu(dx)| dz \\
\le & \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}} |f(z + m_n(x) - m(x) \mid x) - f(z \mid x)| dz \right) \mu(dx) \\
= & \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}} |f(z + m_n(x) - m(x) \mid x) - f(z \mid x)| dz \right) I_{\{|m_n(x) - m(x)| \le \delta\}} \mu(dx) \\
& + \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}} |f(z + m_n(x) - m(x) \mid x) - f(z \mid x)| dz \right) I_{\{|m_n(x) - m(x)| > \delta\}} \mu(dx) \\
\le & \int_{\mathbb{R}^d} \Delta_x(\delta) \mu(dx) + 2\mathbb{P}\{|m(X) - m_n(X)| > \delta \mid (X_1, Y_1), \ldots, (X_n, Y_n)\} \\
= & \int_{\mathbb{R}^d} \Delta_x(\delta) \mu(dx) + 2 \frac{\int_{\mathbb{R}^d} (m(x) - m_n(x))^2 \mu(dx)}{\delta^2} \\
\to & \int_{\mathbb{R}^d} \Delta_x(\delta) \mu(dx)
\end{aligned}$$

a.s. as $n \to \infty$. $\Delta_x(\delta) \le 2$ and for any fixed $x$, $\Delta_x(\delta) \to 0$ as $\delta \to 0$ (cf. the proof of Theorem 2.1 in Devroye and Györfi [8]), therefore the dominated convergence theorem implies that

$$\int_{\mathbb{R}^d} \Delta_x(\delta) \mu(dx) \to 0$$

as $\delta \to 0$, which yields (6). Apply the decomposition

$$f_n(z) - f(z) = V_n(z) + B_n(z),$$

where the variation term is

$$V_n(z) = \frac{1}{n} \sum_{i=1}^{n} \left[ K_{h_i}(z - Z_i) - \mathbb{E}\left\{ K_{h_i}(z - Z_i) \mid (X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}) \right\} \right],$$

while the (conditional) bias term is

$$B_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{ K_{h_i}(z - Z_i) \mid (X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}) \right\} - f(z).$$

Concerning the bias term, $\lim_{n \to \infty} h_n = 0$ and (6) imply that

$$\int_{\mathbb{R}} |B_n(z)| dz = \int_{\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z - u) g_{i-1}(u) du - f(z) \right| dz$$

$$
\begin{aligned}
&\leq \int_{\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) f(u) du - f(z) \right| dz \\
&\quad + \int_{\mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) |g_{i-1}(u) - f(u)| du dz \\
&\leq \int_{\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} K_{h_i}(z-u) f(u) du - f(z) \right| dz + \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} |g_{i-1}(u) - f(u)| du \\
&\to 0
\end{aligned}
$$

a.s., because of the Toeplitz lemma and Theorem 2.1 in Devroye and Györfi [8]. $V_n(\cdot)$ is an average of $L_2$-valued sequence of martingale differences. We apply the generalized Chow theorem [5]: let $U_n$, $n = 1, 2, \ldots$ be an $L_2$-valued sequence of martingale differences such that

$$
\sum_{n=1}^{\infty} \frac{\mathbb{E}\{\|U_n\|_2^2\}}{n^2} < \infty
$$

where $\|\cdot\|_2$ denotes the $L_2$ norm. Then

$$
\lim_{n \to \infty} \left\| \frac{1}{n} \sum_{i=1}^{n} U_i \right\|_2 = 0
$$

a.s. (cf. Györfi, Györfi and Vajda [11]). One has to verify the condition of the generalized Chow theorem:

$$
\begin{aligned}
&\sum_{n=1}^{\infty} \frac{\mathbb{E}\left\{ \|K_{h_n}(\cdot - Z_n) - \mathbb{E}\left\{K_{h_n}(\cdot - Z_n) \mid (X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})\right\}\|_2^2 \right\}}{n^2} \\
\leq &\sum_{n=1}^{\infty} \frac{\mathbb{E}\left\{ \|K_{h_n}(\cdot - Z_n)\|_2^2 \right\}}{n^2} \leq \sum_{n=1}^{\infty} \frac{\|K\|_2^2}{n^2 h_n} < \infty,
\end{aligned}
$$

by the condition of the theorem, and so

$$
\|V_n\|_2 \to 0
$$

a.s. Put

$$
\hat{f}_n(z) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left\{K_{h_i}(z - Z_i) \mid (X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1})\right\}.
$$

then we proved that

$$
\|\hat{f}_n - f\|_1 = \|B_n\|_1 \to 0
$$

a.s., where $\|\cdot\|_1$ denotes the $L_1$ norm, and

$$
\|\hat{f}_n - f_n\|_2 = \|V_n\|_2 \to 0
$$

a.s. From Lemma 3.1 in Györfi, Masry [12] we get that these two limit relations imply

$$
\|f_n - f\|_1 \to 0
$$

a.s.

□

## 3. A non-recursive kernel estimate

Next we introduce a data splitting scheme. Assume that we are given two independent samples:

$$D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

and

$$D'_n = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}.$$

From sample $D_n$ we generate a strongly universally consistent regression estimate $m_n$. Then the non-recursive kernel estimate is defined by

$$f_n(z) := \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(z - Z_i), \tag{7}$$

where in the $i$-th term we plug-in the approximation of the $i$-th residual

$$Z_i := Y'_i - m_n(X'_i).$$

Given $D_n$, the common density of $Z_i$'s is $g_n$.

Under the additive noise model (3), Devroye et al. [7] proved the universal strong consistency of $f_n$ defined by (7).

**Theorem 2.** *Suppose that we are given a strongly universally consistent regression estimate $m_n$, i.e.,*

$$\int_{\mathbb{R}^d} (m(x) - m_n(x))^2 \mu(dx) \to 0 \ \textit{a.s.}$$

*and for given $X = x$, the conditional density of the residual $Y - m(X)$ exists. Assume that the kernel function $K$ is a square integrable density, and*

$$\lim_{n\to\infty} h_n = 0 \ \textit{and} \ \lim_{n\to\infty} nh_n = \infty. \tag{8}$$

*Then*

$$\lim_{n\to\infty} \int_{\mathbb{R}} |f_n(z) - f(z)| dz = 0$$

*a.s.*

Proof. Applying the argument in Devroye [6] we get that

$$\mathbb{P}\left\{ \Big| \int_{\mathbb{R}} |f_n - f| - \mathbb{E}\left\{ \int_{\mathbb{R}} |f_n - f| \mid D_n \right\} \Big| \geq \epsilon \mid D_n \right\} \leq 2e^{-n\epsilon^2/2},$$

therefore one has to prove that

$$\mathbb{E}\left\{ \int_{\mathbb{R}} |f_n - f| \mid D_n \right\} \to 0$$

a.s. Concerning the conditional bias term, we have that

$$\begin{aligned}
& \int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f(z)|dz \\
= & \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} K_{h_n}(z-u) g_n(u) du - f(z) \Big| dz \\
\leq & \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} K_{h_n}(z-u) f(u) du - f(z) \Big| dz + \int_{\mathbb{R}} \int_{\mathbb{R}} K_{h_n}(z-u) |g_n(u) - f(u)| du dz \\
\leq & \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} K_{h_n}(z-u) f(u) du - f(z) \Big| dz + \int_{\mathbb{R}} |g_n(u) - f(u)| du \\
\to & \ 0
\end{aligned}$$

a.s., because of Theorem 2.1 in Devroye, Györfi [8] and (6). For the conditional variation term, let $I$ be an arbitrary interval, then we have that

$$\begin{aligned}
& \mathbb{E}\left\{ \int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f_n(z)| dz \mid D_n \right\} \\
\leq & \int_I \mathbb{E}\left\{ |\mathbb{E}\{f_n(z) \mid D_n\} - f_n(z)| \mid D_n \right\} dz + 2 \int_{I^c} \mathbb{E}\{f_n(z) \mid D_n\} dz \\
\leq & \int_I \sqrt{\mathbb{E}\left\{ |\mathbb{E}\{f_n(z) \mid D_n\} - f_n(z)|^2 \mid D_n \right\}} dz \\
& + 2 \int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f(z)| dz + 2 \int_{I^c} f(z) dz.
\end{aligned}$$

For $\epsilon > 0$, with probability one choose $I$ and $n$ such that

$$2 \int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f(z)| dz + 2 \int_{I^c} f(z) dz < \epsilon.$$

Thus,

$$\begin{aligned}
& \mathbb{E}\left\{ \int_{\mathbb{R}} |\mathbb{E}\{f_n(z) \mid D_n\} - f_n(z)| dz \mid D_n \right\} \\
\leq & \int_I \sqrt{\frac{\mathbb{E}\left\{ |\mathbb{E}\{K_{h_n}(z-Z_1) \mid D_n\} - K_{h_n}(z-Z_1)|^2 \mid D_n \right\}}{n}} dz + \epsilon \\
\leq & \int_I \sqrt{\frac{\mathbb{E}\left\{ K_{h_n}(z-Z_1)^2 \mid D_n \right\}}{n}} dz + \epsilon \\
\leq & \sqrt{\frac{\|K\|_2^2 |I|}{n h_n}} + \epsilon \\
\to & \ \epsilon
\end{aligned}$$

a.s., where $|I|$ denotes the length of the interval $I$. □

**Remark 1.** Using a tricky counter example, Devroye et al. [7] showed that the condition of the existence of conditional densities of the residual cannot be weakened, if for the regression estimate merely strong universal consistency is assumed. The example is follows: Choose $X$ uniformly distributed on $[0, 1]$, let $U$ be independent of $X$ take on

values 1 and $-1$ with probability $1/2$, resp., and set $Y = U \cdot X$. Then $Y$ is uniformly distributed on $[-1, 1]$ and has a density, the regression function is 0. However, $Y = Y - m(X)$ is conditioned on the value of $X = x$ concentrated on $-x$ and $x$ and has no density. Then they constructed an approximation $m_n$ of the regression function such that $\max_x |m_n(x)| \leq \sqrt{h_n} \to 0$ and

$$\liminf_n \int_{\mathbb{R}} |f_n(z) - f(z)| dz \geq 1$$

a.s., where the kernel $K$ is the window kernel.

## References


[1] Ahmad, I. A. Residuals density estimation in nonparametric regression. *Statistics and Probability Letters*, 14, pp. 133-139, 1992.

[2] Akritas, M. G. and Van Keilegom, I. Non-parametric estimation of the residual distribution. *Board of the Foundation of the Scandinavian Journal of Statistics*, Blackwell Publishers Ltd, 28, pp. 549-567, 2001.

[3] Cheng, F. Consistency of error density and distribution function estimators in nonparametric regression. *Statistics and Probability Letters*, 59, pp. 257-270, 2002.

[4] Cheng, F. Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression. *Journal of Statistical Planning and Inference*, 119, pp. 95-107, 2004.

[5] Chow, Y. S. Local convergence of martingales and the law of large numbers. *Annals of Mathematical Statistics*, 36, pp. 552–558, 1965.

[6] Devroye, L. Exponential inequalities in nonparametric estimation. In *Nonparametric Functional Estimation and Related Topics*, G. Roussas (Ed.), NATO ASI Series, Kluwer Academic Publishers, Dordrecht, pp. 31–44, 1991.

[7] Devroye, L., Felber, T., Kohler, M. and Krzyzak, A. $L1$-consistent estimation of the density of residuals in random design regression models. *Statistics and Probability Letters*, 82:173-179, 2012.

[8] Devroye, L. and Györfi, L. *Nonparametric Density Estimation: The $L_1$ View*. John Wiley, New York, 1985.

[9] Efromovich, S. Estimation of the density of regression errors. *Annals of Statistics*, 33, pp. 2194-2227, 2005.

[10] Efromovich, S. Optimal nonparametric estimation of the density of regression errors with finite support. *AISM*, 59, pp. 617-654, 2006.

[11] Györfi, L., Györfi, Z. and Vajda, I. A strong law of large numbers and some applications, *Studia Sci. Math. Hungar.*, 12, pp. 233-244, 1977.

[12] Györfi, L. and Masry, E. The $L_1$ and $L_2$ strong consistency of recursive kernel density estimation from dependent samples, *IEEE Trans. Information Theory*, 36, pp. 531-539, 1990.

[13] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2002.

[14] Neumeyer, N. and Van Keilegom, I. Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, 101, pp. 1067-1078, 2010.

[15] Stout, W. F. *Almost sure convergence*. New York: Academic Press, 1974.

[16] Yamato, H. Sequential estimation of a continuous probability density function and mode. *Bull. Math. Statist.*, 14, pp. 1–12, 1971.

[17] Wolverton, C. T. and Wagner, T. J. Asymptotically optimal discriminant functions for pattern classification. *IEEE Trans. Information Theory*, IT-15, pp. 258–265, 1969.