KÖFOP-2.1.2-VEKOP-15. „A jó kormányzást megalapozó közszolgálat-fejlesztés”

László Györfi

Nonparametric Decisions

Nonparametric Decisions

László Györfi [1]

February 6, 2018

[1]The work was created in commission of the National University of Public Service under the priority project KÖFOP-2.1.2-VEKOP-15-2016-00001 titled "Public Service Development Establishing Good Governance" in Ludovika Research Group.

Contents

Chapter 1

Pattern Recognition

1.1 Bayes decision and its approximation for general loss function

For the statistical pattern recognition, called also classification, a d-dimensional observation vector $\mathbf{X}$ is given, and based on $\mathbf{X}$, the statistician has to make an inference on a random label Y, which takes finitely many values, i.e., it takes values from the set $\{1, 2, \ldots, M\}$. The label Y is called class, too. In fact, the inference is a decision formulated by a decision function

$$g : \mathbb{R}^d \to \{1, 2, \ldots, M\}.$$

If $g(\mathbf{X}) \neq Y$ then the decision makes error.

In the formulation of the Bayes decision problem, introduce a cost function $C(y, y') \geq 0$, which is the cost if the label $Y = y$ and the decision $g(\mathbf{X}) = y'$. For a decision function g, the risk is the expectation of the cost:

$$R(g) = \mathbb{E}\{C(Y, g(\mathbf{X}))\}.$$

In Bayes decision problem, the aim is to minimize the risk, i.e., the goal is to find a function

$$g^* : \mathbb{R}^d \to \{1, 2, \ldots, M\}$$

such that

$$R(g^*) = \min_{g:\mathbb{R}^d \to \{1,2,\ldots,M\}} R(g), \tag{1.1}$$

where g^* is called the Bayes decision function, and $R^* = R(g^*)$ is the Bayes risk.

For the posteriori probabilities, introduce the notations:

$$P_y(\mathbf{X}) = \mathbb{P}\{Y = y \mid \mathbf{X}\}.$$

Let the decision function g^* be defined by

$$g^*(\mathbf{X}) = \arg\min_{y'} \sum_{y=1}^{M} C(y, y') P_y(\mathbf{X}).$$

If $\arg\min$ is not unique then choose the smallest y', which minimizes $\sum_{y=1}^{M} C(y, y') P_y(\mathbf{X})$. This definition implies that for any decision function g,

$$\sum_{y=1}^{M} C(y, g^*(\mathbf{X})) P_y(\mathbf{X}) \leq \sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X}). \tag{1.2}$$

Theorem 1.1. *For any decision function g, we have that*

$$R(g^*) \leq R(g).$$

PROOF. Let $\mathbb{I}$ denote the indicator function. For a decision function g, let's calculate the risk.

$$\begin{aligned}
R(g) &= \mathbb{E}\{C(Y, g(\mathbf{X}))\} \\
&= \mathbb{E}\{\mathbb{E}\{C(Y, g(\mathbf{X})) \mid \mathbf{X}\}\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{M}\sum_{y'=1}^{M} C(y, y') \mathbb{P}\{Y = y, g(\mathbf{X}) = y' \mid \mathbf{X}\}\right\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{M}\sum_{y'=1}^{M} C(y, y') \mathbb{I}_{\{g(\mathbf{X})=y'\}} \mathbb{P}\{Y = y \mid \mathbf{X}\}\right\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X})\right\}.
\end{aligned}$$

(1.2) implies that

$$\begin{aligned}
R(g) &= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X})\right\} \\
&\geq \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g^*(\mathbf{X})) P_y(\mathbf{X})\right\} \\
&= R(g^*).
\end{aligned}$$

□

If the distribution of the observation vector $\mathbf{X}$ has density, then the Bayes decision has an equivalent formulation. Introduce the notations for density f of $\mathbf{X}$ by

$$\mathbb{P}\{\mathbf{X} \in B\} = \int_B f(\mathbf{x}) d\mathbf{x}$$

and for the conditional densities by

$$\mathbb{P}\{\mathbf{X} \in B \mid Y = y\} = \int_B f_y(\mathbf{x}) d\mathbf{x}$$

and for a priori probabilities

$$q_y = \mathbb{P}\{Y = y\},$$

then it is easy to check that

$$P_y(\mathbf{X}) = \mathbb{P}\{Y = y \mid \mathbf{X} = \mathbf{x}\} = \frac{q_y f_y(\mathbf{x})}{f(\mathbf{x})}$$

and therefore

$$\begin{aligned} g^*(\mathbf{x}) &= \arg\min_{y'} \sum_{y=1}^{M} C(y, y') P_y(\mathbf{x}) \\ &= \arg\min_{y'} \sum_{y=1}^{M} C(y, y') \frac{q_y f_y(\mathbf{x})}{f(\mathbf{x})} \\ &= \arg\min_{y'} \sum_{y=1}^{M} C(y, y') q_y f_y(\mathbf{x}). \end{aligned}$$

From the proof of Theorem 1.1 we may derive a formula for the optimal risk:

$$R(g^*) = \mathbb{E}\left\{ \min_{y'} \sum_{y=1}^{M} C(y, y') P_y(\mathbf{X}) \right\}.$$

If $\mathbf{X}$ has density then

$$\begin{aligned} R(g^*) &= \mathbb{E}\left\{\min_{y'}\sum_{y=1}^{M} C(y,y')\frac{q_y f_y(\mathbf{X})}{f(\mathbf{X})}\right\} \\ &= \int_{\mathbb{R}^d}\min_{y'}\sum_{y=1}^{M} C(y,y')\frac{q_y f_y(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathbb{R}^d}\min_{y'}\sum_{y=1}^{M} C(y,y') q_y f_y(\mathbf{x})d\mathbf{x}. \end{aligned}$$

In practice, the posteriori probabilities $\{P_y(\mathbf{X})\}$ are unknown. If we are given some approximations $\{\hat{P}_y(\mathbf{X})\}$, from which one may derive some approximate decision

$$\hat{g}(\mathbf{X}) = \arg\min_{y'}\sum_{y=1}^{M} C(y,y')\hat{P}_y(\mathbf{X})$$

then the question is how well $R(\hat{g})$ approximates $R(g^*)$.

Lemma 1.1. *Put $C_{max} = \max_{y,y'} C(y,y')$, then*

$$0 \le R(\hat{g}) - R(g^*) \le 2C_{max}\sum_{y=1}^{M}\mathbb{E}\left\{|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\}.$$

Proof. We have that

$$\begin{aligned} R(\hat{g}) - R(g^*) &= \mathbb{E}\left\{\sum_{y=1}^{M} C(y,\hat{g}(\mathbf{X}))P_y(\mathbf{X})\right\} - \mathbb{E}\left\{\sum_{y=1}^{M} C(y,g^*(\mathbf{X}))P_y(\mathbf{X})\right\} \\ &= \mathbb{E}\left\{\sum_{y=1}^{M} C(y,\hat{g}(\mathbf{X}))P_y(\mathbf{X}) - \sum_{y=1}^{M} C(y,\hat{g}(\mathbf{X}))\hat{P}_y(\mathbf{X})\right\} \\ &\quad +\mathbb{E}\left\{\sum_{y=1}^{M} C(y,\hat{g}(\mathbf{X}))\hat{P}_y(\mathbf{X}) - \sum_{y=1}^{M} C(y,g^*(\mathbf{X}))\hat{P}_y(\mathbf{X})\right\} \\ &\quad +\mathbb{E}\left\{\sum_{y=1}^{M} C(y,g^*(\mathbf{X}))\hat{P}_y(\mathbf{X}) - \sum_{y=1}^{M} C(y,g^*(\mathbf{X}))P_y(\mathbf{X})\right\}. \end{aligned}$$

The definition of $\hat{g}$ implies that

$$\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))\hat{P}_y(\mathbf{X}) - \sum_{y=1}^{M} C(y, g^*(\mathbf{X}))\hat{P}_y(\mathbf{X}) \leq 0,$$

therefore

$$\begin{aligned} R(\hat{g}) - R(g^*) &\leq \mathbb{E}\left\{\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\} \\ &\quad + \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g^*(\mathbf{X}))|\hat{P}_y(\mathbf{X}) - P_y(\mathbf{X})|\right\} \\ &\leq 2C_{max}\sum_{y=1}^{M} \mathbb{E}\left\{|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\}. \end{aligned}$$

□

1.2 Bayes decision and its approximation for $0-1$ loss

Concerning the cost function, the most frequently studied example is the so called $0-1$ loss:

$$C(y, y') = \begin{cases} 1 & \text{if} \quad y \neq y', \\ 0 & \text{if} \quad y = y'. \end{cases}$$

For the $0-1$ loss, the corresponding risk is the error probability denoted by L:

$$L(g) = \mathbb{E}\{C(Y, g(\mathbf{X}))\} = \mathbb{E}\{\mathbb{I}_{\{Y \neq g(\mathbf{X})\}}\} = \mathbb{P}\{Y \neq g(\mathbf{X})\},$$

and the Bayes decision is of form

$$g^*(\mathbf{X}) = \arg\min_{y'} \sum_{y=1}^{M} C(y, y')P_y(\mathbf{X}) = \arg\min_{y'} \sum_{y \neq y'} P_y(\mathbf{X}) = \arg\max_{y'} P_{y'}(\mathbf{X}),$$

which is called maximum posteriori decision, too.

For the $0-1$ loss, we get that

$$L(g^*) = \mathbb{E}\left\{\min_{y'}(1 - P_{y'}(\mathbf{X}))\right\},$$

which has the form, for densities,

$$L(g^*) = \int_{\mathbb{R}^d} \min_{y'}(f(\mathbf{x}) - q_{y'} f_{y'}(\mathbf{x}))d\mathbf{x} = 1 - \int_{\mathbb{R}^d} \max_{y'} q_{y'} f_{y'}(\mathbf{x})d\mathbf{x}.$$

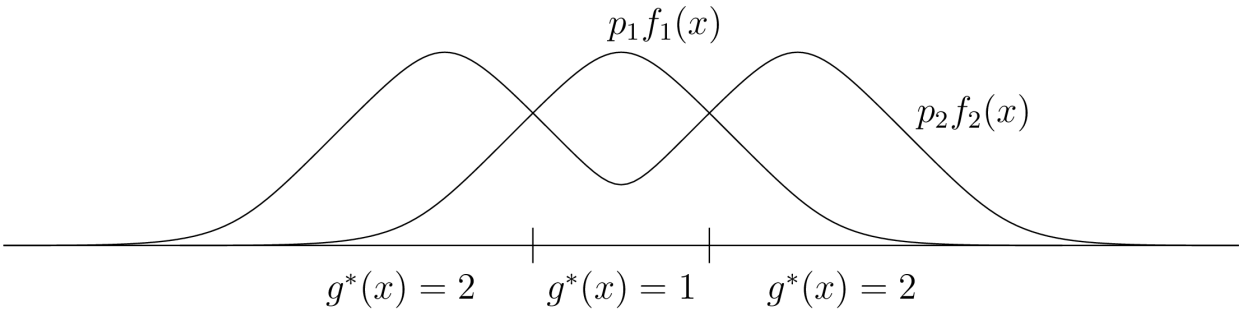

Figure 1.1: Bayes decision.

For the binary classification, $M = 2$, and we have that

$$L(g^*) = \mathbb{E}\left\{\min(P_1(\mathbf{X}), P_2(\mathbf{X}))\right\},$$

and, for densities,

$$L(g^*) = \int_{\mathbb{R}^d} \min(q_1 f_1(\mathbf{x}), q_2 f_2(\mathbf{x}))d\mathbf{x}.$$

Figure 1.1 illustrates the Bayes decision, while the red area in Figure 1.2 is equal to the Bayes error probability.

In the special case of the approximate maximum posteriori decision the inequality in Lemma 1.1 can be slightly improved such that the factor 2 is missing:

$$0 \leq L(\hat{g}) - L(g^*) \leq \sum_{y=1}^{M} \mathbb{E}\left\{|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\}.$$

In the sequel, we study only the case of binary classification. Its extension to the multi-class case is obvious. For the sake of simplicity assume that Y takes values ± 1. Put

$$D(\mathbf{X}) = \mathbb{E}\{Y \mid \mathbf{X}\} = P_1(\mathbf{X}) - P_{-1}(\mathbf{X}).$$

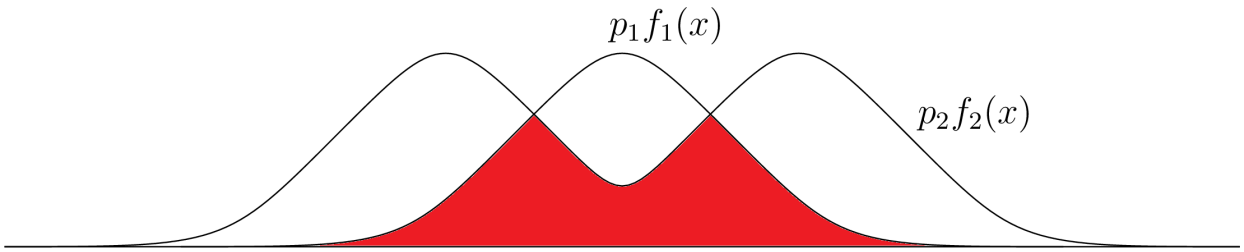

Figure 1.2: Bayes error probability.

Let $sign(x) = 1$ for $x > 0$ and $sign(x) = -1$ for $x \leq 0$. Then the Bayes decision has the form

$$g^*(\mathbf{X}) = sign\, D(\mathbf{X}).$$

For an arbitrary function $\hat{D}$, the corresponding plug-in decision g is defined by

$$g(\mathbf{X}) = sign\, \hat{D}(\mathbf{X}).$$

Theorem 1.2. *For any function $\hat{D}$, we have that*

$$L(g) - L(g^*) = \mathbb{E}\left\{\mathbb{I}_{\{sign\, \hat{D}(\mathbf{X}) \neq sign\, D(\mathbf{X})\}} |D(\mathbf{X})|\right\} \tag{1.3}$$

$$\leq \mathbb{E}\left\{|\hat{D}(\mathbf{X}) - D(\mathbf{X})|\right\} \tag{1.4}$$

$$\leq \sqrt{\mathbb{E}\left\{|\hat{D}(\mathbf{X}) - D(\mathbf{X})|^2\right\}}. \tag{1.5}$$

Proof. The identities

$$P_1(\mathbf{X}) = (1 + D(\mathbf{X}))/2$$

and

$$P_{-1}(\mathbf{X}) = (1 - D(\mathbf{X}))/2$$

imply that

$$\begin{aligned} L(g) &= \mathbb{P}\{Y = 1, g(\mathbf{X}) = -1\} + \mathbb{P}\{Y = -1, g(\mathbf{X}) = 1\} \\ &= \mathbb{E}\{P_1(\mathbf{X})\mathbb{I}_{g(\mathbf{X})=-1}\} + \mathbb{E}\{P_{-1}(\mathbf{X})\mathbb{I}_{g(\mathbf{X})=1}\} \\ &= \mathbb{E}\{(1 + D(\mathbf{X}))/2\mathbb{I}_{g(\mathbf{X})=-1}\} + \mathbb{E}\{(1 - D(\mathbf{X}))/2\mathbb{I}_{g(\mathbf{X})=1}\}. \end{aligned}$$

Thus

$$L(g) - L(g^*) = (sign\, D(\mathbf{X}) - sign\, \hat{D}(\mathbf{X}))D(\mathbf{X})/2,$$

from which we get (1.3). (1.4) is obvious, while (1.5) is just the Cauchy-Schwartz inequality. □

Based on these relations, one can introduce efficient pattern recognition rules. Given data

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\},$$

estimates D_n of the regression function D can be constructed. Then the corresponding plug-in classifier is defined by

$$g_n(\mathbf{x}) = sign\, D_n(\mathbf{x}). \tag{1.6}$$

If the estimate D_n is close to the regression function D, then the error of the plug-in classifier is close to the optimal error. In the next sections some examples are shown for plug-in classifiers.

1.3 The regression problem

In regression analysis one considers a random vector $(\mathbf{X}, Y)$, where $\mathbf{X}$ is $\mathbb{R}^d$-valued and Y is $\mathbb{R}$-valued, and one is interested how the value of the so-called response variable Y depends on the value of the observation vector $\mathbf{X}$. This means that one wants to find a function $h : \mathbb{R}^d \to \mathbb{R}$, such that $h(\mathbf{X})$ is a "good approximation of Y," that is, $h(\mathbf{X})$ should be close to Y in some sense, which is equivalent to making $|h(\mathbf{X}) - Y|$ "small." Since $\mathbf{X}$ and Y are random vectors, $|h(\mathbf{X}) - Y|$ is random as well, therefore it is not clear what "small $|h(\mathbf{X}) - Y|$" means. We can resolve this problem by introducing the so-called L_2 *risk* or *mean squared error* of h,

$$\mathbb{E}|h(\mathbf{X}) - Y|^2,$$

and requiring it to be as small as possible.

So we are interested in a function $m^* : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathbb{E}|m^*(\mathbf{X}) - Y|^2 = \min_{h:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}|h(\mathbf{X}) - Y|^2.$$

Such a function can be obtained explicitly as follows. Let

$$m(\mathbf{x}) = \mathbb{E}\{Y|\mathbf{X} = \mathbf{x}\}$$

be the *regression function*. We will show that the regression function minimizes the L_2 risk. Indeed, for an arbitrary $h : \mathbb{R}^d \to \mathbb{R}$, one has

$$\begin{aligned}\mathbb{E}|h(\mathbf{X}) - Y|^2 &= \mathbb{E}|h(\mathbf{X}) - m(\mathbf{X}) + m(\mathbf{X}) - Y|^2 \\ &= \mathbb{E}|h(\mathbf{X}) - m(\mathbf{X})|^2 + \mathbb{E}|m(\mathbf{X}) - Y|^2,\end{aligned}$$

where we have used

$$\begin{aligned}&\mathbb{E}\left\{(h(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - Y)\right\} \\ &= \mathbb{E}\left\{\mathbb{E}\left\{(h(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - Y)\middle|\mathbf{X}\right\}\right\} \\ &= \mathbb{E}\left\{(h(\mathbf{X}) - m(\mathbf{X}))\mathbb{E}\{m(\mathbf{X}) - Y|\mathbf{X}\}\right\} \\ &= \mathbb{E}\left\{(h(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - m(\mathbf{X}))\right\} \\ &= 0.\end{aligned}$$

Hence,

$$\mathbb{E}|h(\mathbf{X}) - Y|^2 = \int_{\mathbb{R}^d} |h(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) + \mathbb{E}|m(\mathbf{X}) - Y|^2, \tag{1.7}$$

where μ denotes the distribution of $\mathbf{X}$. The first term is called the L_2 error of h. It is always nonnegative and is zero if $h(\mathbf{x}) = m(\mathbf{x})$. Therefore,

$$m^*(\mathbf{x}) = m(\mathbf{x}),$$

i.e., the optimal approximation (with respect to the L_2 risk) of Y by a function of $\mathbf{X}$ is given by $m(\mathbf{X})$.

Denote by $(\mathbf{X}, Y)$, $(\mathbf{X}_1, Y_1)$, $(\mathbf{X}_2, Y_2), \ldots$ independent and identically distributed (i.i.d.) random variables with $\mathbb{E}Y^2 < \infty$. Let $\mathcal{D}_n$ be the set of *data* defined by

$$\mathcal{D}_n = \left\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\right\}.$$

In general, estimates will not be equal to the regression function. To compare different estimates, we need an error criterion which measures the difference between the regression function and an arbitrary estimate m_n. One of the key points we would like to make is that the motivation for introducing the regression function leads naturally to an L_2 error criterion for measuring the performance of the regression function estimate. Recall that the main goal was to find a function h such that the L_2 risk $\mathbb{E}|h(\mathbf{X}) - Y|^2$ is small. The minimal value of this L_2 risk is $\mathbb{E}|m(\mathbf{X}) - Y|^2$, and it is

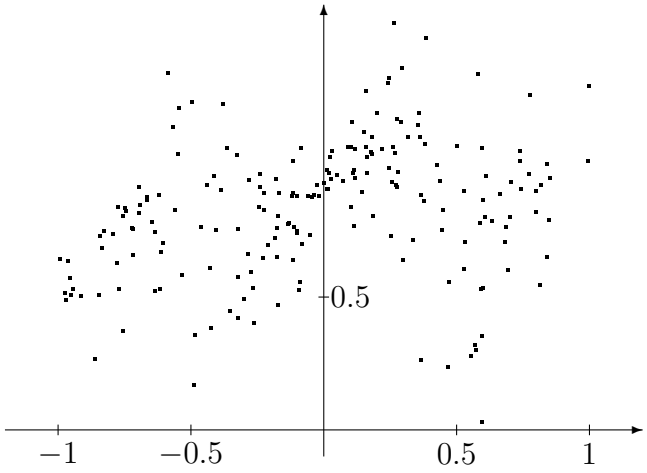

Figure 1.3: Simulated data points.

achieved by the regression function m. Similarly to (1.7), one can show that the L_2 risk $\mathbb{E}\{|m_n(\mathbf{X}) - Y|^2|\mathcal{D}_n\}$ of an estimate m_n satisfies

$$\mathbb{E}\left\{|m_n(\mathbf{X}) - Y|^2|\mathcal{D}_n\right\} = \int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) + \mathbb{E}|m(\mathbf{X}) - Y|^2. \tag{1.8}$$

Thus the L_2 risk of an estimate m_n is close to the optimal value if and only if the L_2 error

$$\int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) \tag{1.9}$$

is close to zero. Therefore we will use the L_2 error (1.9) in order to measure the quality of an estimate and we will study estimates for which this L_2 error is small.

For univariate $X = \mathbf{X}$ one can often use a plot of the simulated data. These data will be used throughout the chapter. They consist of $n = 200$ points such that X is standard normal restricted to $[-1, 1]$, i.e., the density of X is proportional to the standard normal density on $[-1, 1]$ and is zero elsewhere. The regression function is piecewise polynomial:

$$m(x) = \begin{cases} (x+2)^2/2 & \text{if} \quad -1 \le x < -0.5, \\ x/2 + 0.875 & \text{if} \quad -0.5 \le x < 0, \\ -5(x-0.2)^2 + 1.075 & \text{if} \quad 0 < x \le 0.5, \\ x + 0.125 & \text{if} \quad 0.5 \le x < 1. \end{cases}$$

Given X, the conditional distribution of $Y - m(X)$ is normal with mean zero and standard deviation

$$\sigma(X) = 0.2 - 0.1\cos(2\pi X).$$

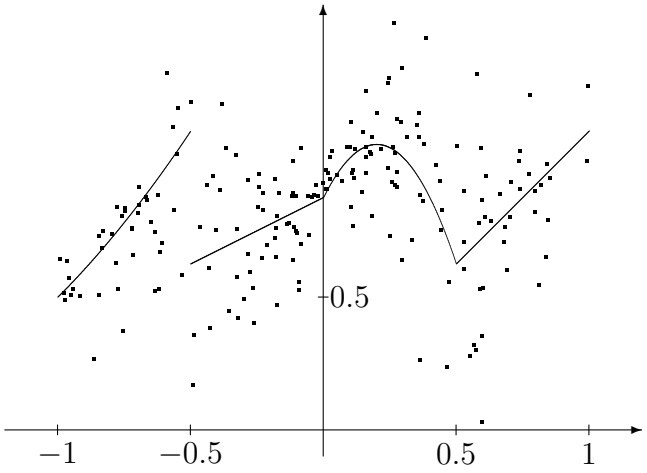

Figure 1.4: Data points and regression function.

Figure 1.3 shows the data points. In this example the human eye is not able to see from the data points what the regression function looks like. In Figure 1.4 the data points are shown together with the regression function.

1.4 The margin condition

Given the plug-in classification rule g_n derived from the regression estimate D_n it follows

$$\mathbb{E}\{L(g_n)\} - L(g^*) \leq \mathbb{E}\{|D(\mathbf{X}) - D_n(\mathbf{X})|\}$$

(cf. Theorem 1.2). Therefore we may get an upper bound on the rate of convergence of the excess error probability $\mathbb{E}\{L(g_n)\} - L(g^*)$ via the L_1 rate of convergence of the corresponding regression estimation.

However, according to Section 6.7 in Devroye, Györfi, Lugosi (1996), the classification is easier than L_1 regression function estimation, since the rate of convergence of the error probability depends on the behavior of the function D in the neighborhood of the decision boundary

$$B_0 = \{\mathbf{x}; D(\mathbf{x}) = 0\}. \tag{1.10}$$

This phenomenon has been discovered by Mammen and Tsybakov (1999), Tsybakov (2004), who formulated the (strong) margin condition:

- *The strong margin condition.* Assume that for all $0 < t \leq 1$,

$$\mathbb{P}\left\{|D(\mathbf{X})| \leq t\right\} \leq ct^{\alpha}, \tag{1.11}$$

  where $\alpha > 0$ and $c > 0$.

Kohler and Krzyżak (2007) introduced the weak margin condition:

- *The weak margin condition.* Assume that for all $0 < t \leq 1$,

$$\mathbb{E}\left\{\mathbb{I}_{\{|D(\mathbf{X})|\leq t\}}|D(\mathbf{X})|\right\} \leq ct^{1+\alpha}. \tag{1.12}$$

Obviously, the strong margin condition implies the weak margin condition:

$$\mathbb{E}\left\{\mathbb{I}_{\{|D(\mathbf{X})|\leq t\}}|D(\mathbf{X})|\right\} \leq \mathbb{E}\left\{\mathbb{I}_{\{|D(\mathbf{X})|\leq t\}}t\right\} = t\mathbb{P}\left\{|D(\mathbf{X})| \leq t\right\} \leq ct \cdot t^{\alpha}.$$

The difference between the strong and weak margin condition is that, for the strong margin condition, the event

$$\{D(\mathbf{X}) = 0\}$$

counts. One can weaken the strong margin condition (1.11) such that we require

$$\mathbb{P}\left\{0 < |D(\mathbf{X})| \leq t\right\} \leq ct^{\alpha}. \tag{1.13}$$

Obviously, (1.13) implies (1.12). The margin conditions measure how fast the probability of a t-neighborhood of the decision boundary increases with t. A large value of α corresponds to a small probability of the neighborhood of the decision boundary, which means that the probability for events far away of the decision boundary is high. Therefore, a classifier can make the right decision more easily, hence one can expect smaller errors for larger values of α.

Audibert and Tsybakov (2005) proved that if the plug-in classifier g has been derived from the regression estimate $\tilde{D}$ and if D satisfies the strong margin condition, then

$$L(g) - L^* \leq \left(\int (\tilde{D}(\mathbf{x}) - D(\mathbf{x}))^2 \mu(d\mathbf{x})\right)^{\frac{1+\alpha}{2+\alpha}}. \tag{1.14}$$

It is easy to see that (1.14) holds even under weak margin condition: (1.3) implies that

$$L(g) - L^* = \mathbb{E}\left\{\mathbb{I}_{\{g(\mathbf{X})\neq g^*(\mathbf{X})\}}|D(\mathbf{X})|\right\}. \tag{1.15}$$

For fixed $t_n > 0$,

$$\begin{aligned} L(g) - L^* &= \mathbb{E}\left\{\mathbb{I}_{\{sign\,\tilde{D}(\mathbf{X}) \neq sign\,D(\mathbf{X}), |D(\mathbf{X})| \leq t_n\}} |D(\mathbf{X})|\right\} \\ &\quad + \mathbb{E}\left\{\mathbb{I}_{\{sign\,\tilde{D}(\mathbf{X}) \neq sign\,D(\mathbf{X}), |D(\mathbf{X})| > t_n\}} |D(\mathbf{X})|\right\} \\ &\leq \mathbb{E}\left\{\mathbb{I}_{\{|D(\mathbf{X})| \leq t_n\}} |D(\mathbf{X})|\right\} \\ &\quad + \mathbb{E}\left\{\mathbb{I}_{\{sign\,\tilde{D}(\mathbf{X}) \neq sign\,D(\mathbf{X}), |\tilde{D}(\mathbf{X}) - D(\mathbf{X})| > t_n\}} |\tilde{D}(\mathbf{X}) - D(\mathbf{X})|\right\}, \end{aligned}$$

therefore the weak margin condition implies that

$$\begin{aligned} L(g) - L^* &\leq c t_n^{1+\alpha} + t_n \mathbb{E}\left\{\mathbb{I}_{\{|\tilde{D}(\mathbf{X}) - D(\mathbf{X})| > t_n\}} \frac{|\tilde{D}(\mathbf{X}) - D(\mathbf{X})|}{t_n}\right\} \\ &\leq c t_n^{1+\alpha} + t_n \mathbb{E}\left\{\frac{|\tilde{D}(\mathbf{X}) - D(\mathbf{X})|^2}{t_n^2}\right\}. \end{aligned}$$

For the choice

$$t_n = \left(\mathbb{E}\left\{|\tilde{D}(\mathbf{X}) - D(\mathbf{X})|^2\right\}\right)^{\frac{1}{2+\alpha}}$$

we get (1.14).

For bounding the error probability, assume, for example, that D satisfies the *Lipschitz condition*: for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$

$$|D(\mathbf{x}) - D(\mathbf{z})| \leq C\|\mathbf{x} - \mathbf{z}\|.$$

If D is Lipschitz continuous and $\mathbf{X}$ is bounded then there are regression estimates such that

$$\mathbb{E}\int (D_n(\mathbf{x}) - D(\mathbf{x}))^2 \mu(dx) \leq c_1^2 n^{-\frac{2}{d+2}},$$

therefore (1.14) means that

$$\mathbb{E}L(g_n) - L^* \leq \left(c_1^2 n^{-\frac{2}{d+2}}\right)^{\frac{1+\alpha}{2+\alpha}} = \left(c_1^{1+\alpha} n^{-\frac{1+\alpha}{d+2}}\right)^{\frac{2}{2+\alpha}}.$$

In the analysis one usually assumes some conditions on the density:

- The *strong density condition* means that for $f(x) > 0$,

$$f(x) \geq f_{min} > 0.$$

- The *weak density condition* means that there exist $c_{min} > 0$ and $\delta > 0$ such that for $r \leq \delta$,

$$\mu(S_{x,r}) \geq c_{min}^d f(x) r^d.$$

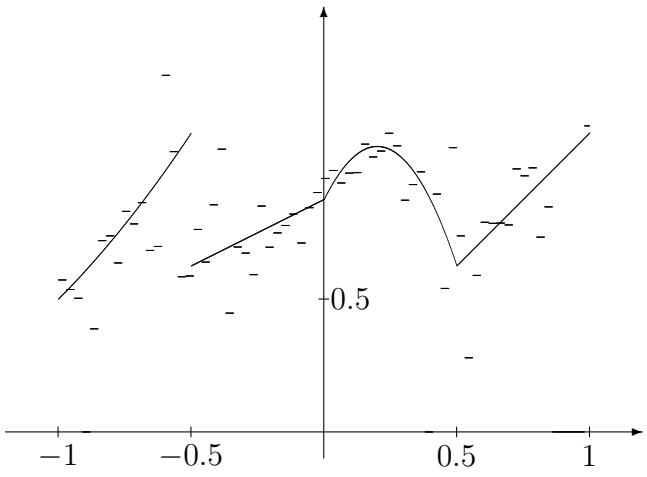

Figure 1.5: Undersmoothing: $h = 0.03$, L_2 error $= 0.062433$.

1.5 Partitioning classifier

Partitioning regression estimate

Let $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots\}$ be a partition of $\mathbb{R}^d$ and for each $\mathbf{x} \in \mathbb{R}^d$ let $A_n(\mathbf{x})$ denote the cell of $\mathcal{P}_n$ containing $\mathbf{x}$. The partitioning estimate (histogram) of the regression function is defined as

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}$$

with $0/0 = 0$ by definition. This means that the partitioning estimate is a local averaging estimate such for a given $\mathbf{x}$ we take the average of those Y_i's for which $\mathbf{X}_i$ belongs to the same cell into which $\mathbf{x}$ falls.

The simplest version of this estimate is obtained for $d = 1$ and when the cells $A_{n,j}$ are intervals of size $h = h_n$. Figures 1.5 – 1.7 show the estimates for various choices of h for our simulated data. In the first figure h is too small (undersmoothing, large variance), in the second choice it is about right, while in the third it is too large (oversmoothing, large bias).

For $d > 1$ one can use, e.g., a cubic partition, where the cells $A_{n,j}$ are cubes of volume h_n^d, or a rectangle partition which consists of rectangles $A_{n,j}$ with side lengths $h_{n1}, \ldots, h_{nd}$. For the sake of illustration we generated two-dimensional data when the actual distribution is a correlated normal distribution. The partition in Figure 1.8 is cubic, and the partition in Figure 1.9 is made of rectangles.

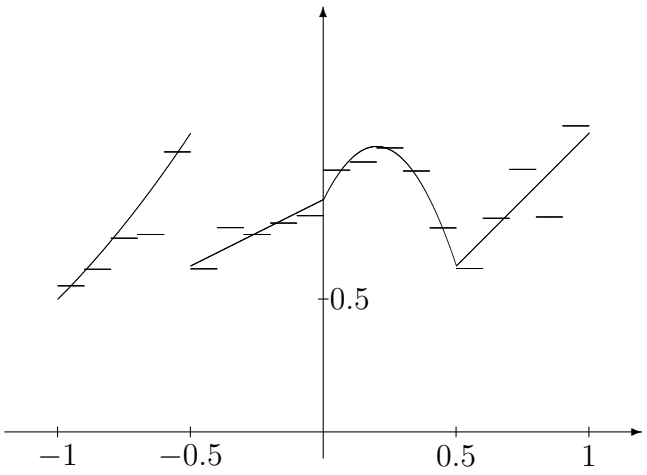

Figure 1.6: Good choice: $h = 0.1$, L_2 error = 0.003642.

Cubic and rectangle partitions are particularly attractive from the computational point of view, because the set $A_n(\mathbf{x})$ can be determined for each $\mathbf{x}$ in constant time, provided that we use an appropriate data structure. In most cases, partitioning estimates are computationally superior to the other nonparametric estimates, particularly if the search for $A_n(\mathbf{x})$ is organized using binary decision trees (cf. Friedman (1977)).

Another advantage of the partitioning estimate is that it can be represented or com-

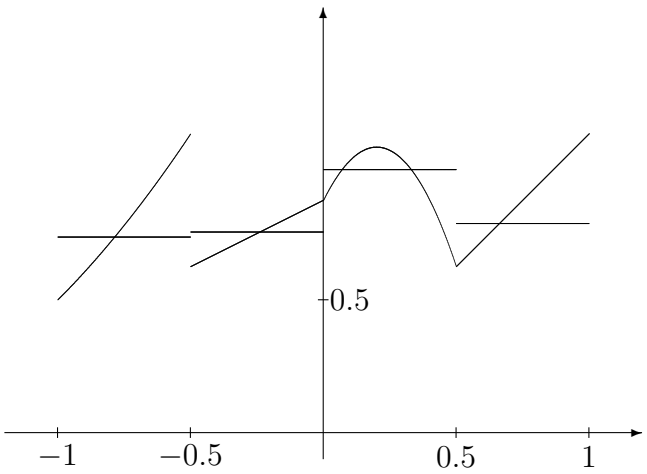

Figure 1.7: Oversmoothing: $h = 0.5$, L_2 error = 0.013208.

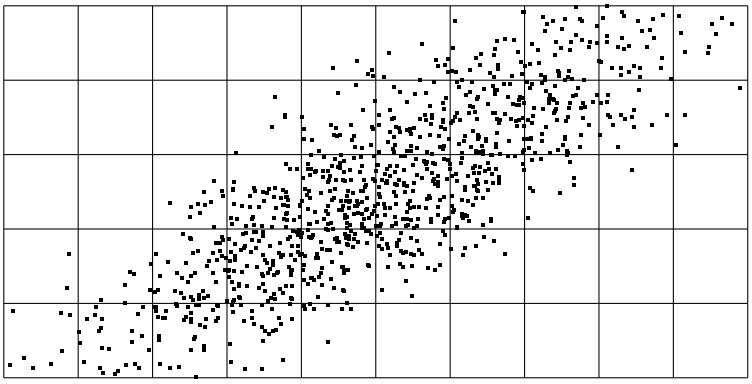

Figure 1.8: Cubic partition.

pressed very efficiently. Instead of storing all data D_n, one should only know the estimate for each nonempty cell, i.e., for cells $A_{n,j}$ for which $\mu_n(A_{n,j}) > 0$, where μ_n denotes the empirical distribution. The number of nonempty cells is much smaller than n. (Cf. Lugosi, Nobel (1996).)

Inequalities for independent random variables

Next we summarize some inequalities for the sum of independent random variables, which are used in the analysis of classification error probabilities.

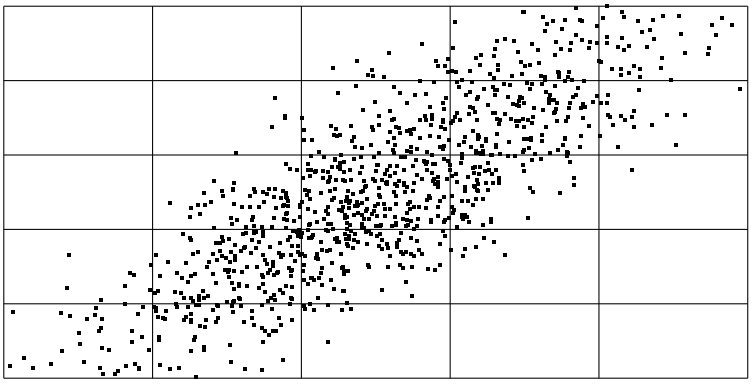

Figure 1.9: Rectangle partition.

Lemma 1.2. (CHERNOFF (1952)). *Let B be a binomial random variable with parameters n and p. Then, for $1 > \epsilon > p > 0$,*

$$\mathbb{P}\{B > n\epsilon\} \leq e^{-n\left[\epsilon \log \frac{\epsilon}{p} + (1-\epsilon)\log \frac{1-\epsilon}{1-p}\right]} \leq e^{-n[p-\epsilon+\epsilon\log(\epsilon/p)]}$$

and, for $0 < \epsilon < p < 1$,

$$\mathbb{P}\{B < n\epsilon\} \leq e^{-n\left[\epsilon \log \frac{\epsilon}{p} + (1-\epsilon)\log \frac{1-\epsilon}{1-p}\right]} \leq e^{-n[p-\epsilon+\epsilon\log(\epsilon/p)]}.$$

PROOF. We proceed by Chernoff's exponential bounding method. In particular, for arbitrary $s > 0$,

$$\begin{aligned}
\mathbb{P}\{B > n\epsilon\} &= \mathbb{P}\{sB > sn\epsilon\} \\
&= \mathbb{P}\{e^{sB} > e^{sn\epsilon}\} \\
&\leq e^{-sn\epsilon}\mathbb{E}\{e^{sB}\} \\
&\quad \text{(by the Markov inequality)} \\
&= e^{-sn\epsilon}\sum_{k=0}^{n} e^{sk}\binom{n}{k}p^k(1-p)^{n-k} \\
&= e^{-sn\epsilon}(e^s p + 1 - p)^n \\
&= [e^{-s\epsilon}(e^s p + 1 - p)]^n.
\end{aligned}$$

Next choose s such that

$$e^s = \frac{\epsilon}{1-\epsilon}\frac{1-p}{p}.$$

With this value we get

$$\begin{aligned}
e^{-s\epsilon}(e^s p + 1 - p) &= e^{-\epsilon\cdot\log\left(\frac{\epsilon}{1-\epsilon}\frac{1-p}{p}\right)} \cdot \left(\frac{\epsilon}{1-\epsilon}\frac{1-p}{p}\cdot p + 1 - p\right) \\
&= e^{-\epsilon\cdot\log\left(\frac{\epsilon}{p}\frac{1-p}{1-\epsilon}\right)} \cdot \left(\epsilon\cdot\frac{1-p}{1-\epsilon} + 1 - p\right) \\
&= e^{-\epsilon\cdot\log\frac{\epsilon}{p} - \epsilon\cdot\log\frac{1-p}{1-\epsilon} + \log\frac{1-p}{1-\epsilon}} \\
&= e^{-\epsilon\cdot\log\frac{\epsilon}{p} + (1-\epsilon)\cdot\log\frac{1-p}{1-\epsilon}},
\end{aligned}$$

which implies the first inequality.

The second inequality follows from

$$\begin{aligned}(1-\epsilon)\log\frac{1-\epsilon}{1-p} &= -(1-\epsilon)\log\frac{1-p}{1-\epsilon}\\ &= -(1-\epsilon)\log\left(1+\frac{\epsilon-p}{1-\epsilon}\right)\\ &\geq -(1-\epsilon)\cdot\frac{\epsilon-p}{1-\epsilon}\\ &\quad (\text{by } \log(1+x)\leq x)\\ &= p-\epsilon.\end{aligned}$$

To prove the second half of the lemma, observe that $n-B$ is a binomial random variable with parameters n and $1-p$. Hence for $\epsilon < p$ the results of the first step imply that

$$\begin{aligned}\mathbb{P}\{B < n\epsilon\} &= \mathbb{P}\{n-B > n(1-\epsilon)\}\\ &\leq e^{-n\left[(1-\epsilon)\log\frac{1-\epsilon}{1-p}+\epsilon\log\frac{\epsilon}{p}\right]}\\ &= e^{-n\left[\epsilon\log\frac{\epsilon}{p}+(1-\epsilon)\log\frac{1-\epsilon}{1-p}\right]}\\ &\leq e^{-n[p-\epsilon+\epsilon\log(\epsilon/p)]}.\end{aligned}$$

□

Lemma 1.3. (Bernstein (1946)). *Let $X_1,\ldots,X_n$ be independent real-valued random variables, let $a,b\in\mathbb{R}$ with $a<b$, and assume that $X_i\in[a,b]$ with probability one $(i=1,\ldots,n)$. Let*

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}\mathbb{V}ar\{X_i\}>0.$$

Then, for all $\epsilon>0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mathbb{E}\{X_i\})\right|>\epsilon\right\}\leq 2e^{-\frac{n\epsilon^2}{2\sigma^2+2\epsilon(b-a)/3}}.$$

Proof. Set $Y_i=X_i-\mathbb{E}\{X_i\}$ $(i=1,\ldots,n)$. Then we have, with probability one,

$$|Y_i|\leq b-a \quad \text{and} \quad \mathbb{E}\{Y_i^2\}=\mathbb{V}ar\{X_i\} \quad (i=1,\ldots,n).$$

By Chernoff's exponential bounding method we get, for arbitrary $s > 0$,

$$\begin{aligned}
\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}\{X_i\}) > \epsilon\right\} &= \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i > \epsilon\right\} \\
&= \mathbb{P}\left\{s\sum_{i=1}^{n} Y_i - sn\epsilon > 0\right\} \\
&\leq \mathbb{E}\left\{e^{s\sum_{i=1}^{n} Y_i - sn\epsilon}\right\} \\
&= e^{-sn\epsilon}\prod_{i=1}^{n}\mathbb{E}\{e^{sY_i}\},
\end{aligned}$$

by the independence of Y_i's. Because of $|Y_i| \leq b - a$ a.s.

$$\begin{aligned}
e^{sY_i} &= 1 + sY_i + \sum_{j=2}^{\infty}\frac{(sY_i)^j}{j!} \\
&\leq 1 + sY_i + \sum_{j=2}^{\infty}\frac{s^j Y_i^2 (b-a)^{j-2}}{2\cdot 3^{j-2}} \\
&= 1 + sY_i + \frac{s^2 Y_i^2}{2}\sum_{j=2}^{\infty}\left(\frac{s\,(b-a)}{3}\right)^{j-2} \\
&= 1 + sY_i + \frac{s^2 Y_i^2}{2}\frac{1}{1 - s(b-a)/3}
\end{aligned}$$

if $|s(b-a)/3| < 1$. This, together with $\mathbb{E}\{Y_i\} = 0$ $(i = 1, \ldots, n)$ and $1 + x \leq e^x$ $(x \in \mathbb{R})$, implies

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}\{X_i\}) > \epsilon\right\} \\
\leq\ & e^{-sn\epsilon}\prod_{i=1}^{n}\left(1 + \frac{s^2\,\mathbb{V}ar\{X_i\}}{2}\frac{1}{1 - s(b-a)/3}\right) \\
\leq\ & e^{-sn\epsilon}\prod_{i=1}^{n}\exp\left(\frac{s^2\,\mathbb{V}ar\{X_i\}}{2}\frac{1}{1 - s(b-a)/3}\right) \\
=\ & \exp\left(-sn\epsilon + \frac{s^2 n\sigma^2}{2(1 - s(b-a)/3)}\right).
\end{aligned}$$

Set

$$s = \frac{\epsilon}{\epsilon(b-a)/3 + \sigma^2}.$$

Then

$$\left| \frac{s(b-a)}{3} \right| < 1$$

and

$$\begin{aligned}
&-sn\epsilon + \frac{s^2 n\sigma^2}{2(1 - s(b-a)/3)} \\
&= \frac{-n\epsilon^2}{\epsilon(b-a)/3 + \sigma^2} + \frac{\epsilon^2}{(\epsilon(b-a)/3 + \sigma^2)^2} \cdot \frac{n\sigma^2}{2\left(1 - \frac{\epsilon(b-a)/3}{\epsilon(b-a)/3 + \sigma^2}\right)} \\
&= \frac{-n\epsilon^2}{\epsilon(b-a)/3 + \sigma^2} + \frac{\epsilon^2}{\epsilon(b-a)/3 + \sigma^2} \cdot \frac{n\sigma^2}{2\left(\epsilon(b-a)/3 + \sigma^2 - \epsilon(b-a)/3\right)} \\
&= \frac{-n\epsilon^2}{2\epsilon(b-a)/3 + 2\sigma^2},
\end{aligned}$$

hence

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) > \epsilon \right\} \leq \exp\left(\frac{-n\epsilon^2}{2\epsilon(b-a)/3 + 2\sigma^2} \right).$$

Similarly,

$$\begin{aligned}
\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) < -\epsilon \right\} &= \mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^{n} (-X_i - \mathbb{E}\{-X_i\}) > \epsilon \right\} \\
&\leq \exp\left(\frac{-n\epsilon^2}{2\epsilon(b-a)/3 + 2\sigma^2} \right),
\end{aligned}$$

which implies the assertion. □

Lemma 1.4. (HOEFFDING (1963)). *Let $X_1, \ldots, X_n$ be independent real-valued random variables, let $a_1, b_1, \ldots, a_n, b_n \in \mathbb{R}$, and assume that $X_i \in [a_i, b_i]$ with probability one $(i = 1, \ldots, n)$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}\{X_i\}) \right| > \epsilon \right\} \leq 2e^{-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n} |b_i - a_i|^2}}.$$

PROOF. Let $s > 0$ be arbitrary. Similarly to the proof of Lemma 1.3 we get

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) > \epsilon\right\} \\
\leq\ & \exp(-sn\epsilon)\cdot\prod_{i=1}^{n}\mathbb{E}\left\{\exp\left(s\cdot(X_i - \mathbb{E}X_i)\right)\right\}.
\end{aligned}$$

We will show momentarily

$$\mathbb{E}\left\{\exp\left(s\cdot(X_i - \mathbb{E}X_i)\right)\right\} \leq \exp\left(\frac{s^2(b_i-a_i)^2}{8}\right) \quad (i = 1,\ldots,n), \tag{1.16}$$

from which we can conclude

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) > \epsilon\right\} \leq \exp\left(-sn\epsilon + \frac{s^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2\right).$$

The right-hand side is minimal for

$$s = \frac{4\,n\,\epsilon}{\sum_{i=1}^{n}(b_i-a_i)^2}.$$

With this value we get

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) > \epsilon\right\} \\
\leq\ & \exp\left(-\frac{4n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i-a_i)^2} + \frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i-a_i)^2}\right) \\
=\ & \exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i-a_i)^2}\right).
\end{aligned}$$

This implies that

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| > \epsilon\right\} \\
&= \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) > \epsilon\right\} + \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}(-X_i - \mathbb{E}\{-X_i\}) > \epsilon\right\} \\
&\leq 2\exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i-a_i)^2}\right).
\end{aligned}$$

So it remains to show (1.16). Fix $i \in \{1, \ldots, n\}$ and set

$$Y = X_i - \mathbb{E}X_i.$$

Then $Y \in [a_i - \mathbb{E}X_i, b_i - \mathbb{E}X_i] =: [a, b]$ with probability one, $a - b = a_i - b_i$, and $\mathbb{E}Y = 0$. We have to show

$$\mathbb{E}\left\{\exp(sY)\right\} \leq \exp\left(\frac{s^2(b-a)^2}{8}\right). \tag{1.17}$$

Because of e^{sx} convex we have

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{for all } a \leq x \leq b,$$

thus

$$\begin{aligned}
\mathbb{E}\{\exp(sY)\} &\leq \frac{\mathbb{E}\{Y\} - a}{b-a}e^{sb} + \frac{b - \mathbb{E}\{Y\}}{b-a}e^{sa} \\
&= e^{sa}\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{s(b-a)}\right) \\
&\qquad \text{(because of } \mathbb{E}\{Y\} = 0\text{)}.
\end{aligned}$$

Setting

$$p = -\frac{a}{b-a}$$

we get

$$\mathbb{E}\{\exp(sY)\} \leq (1 - p + p \cdot e^{s(b-a)})e^{-s\,p\,(b-a)} = e^{\Phi(s(b-a))},$$

where

$$\Phi(u) = \ln\left((1 - p + pe^u)e^{-pu}\right) = \ln\left(1 - p + pe^u\right) - pu.$$

Next we make a Taylor expansion of Φ. Because of

$$\Phi(0) = 0,$$

$$\Phi'(u) = \frac{pe^u}{1 - p + pe^u} - p, \text{ hence } \Phi'(0) = 0$$

and

$$\begin{aligned}
\Phi''(u) &= \frac{(1 - p + pe^u)pe^u - pe^u pe^u}{(1 - p + pe^u)^2} = \frac{(1-p)pe^u}{(1 - p + pe^u)^2} \\
&\leq \frac{(1-p)pe^u}{4(1-p)pe^u} = \frac{1}{4}
\end{aligned}$$

we get, for any $u > 0$,

$$\Phi(u) = \Phi(0) + \Phi'(0)u + \frac{1}{2}\Phi''(\eta)u^2 \leq \frac{1}{8}u^2$$

for some $\eta \in [0, u]$. We conclude

$$\mathbb{E}\{\exp(sY)\} \leq e^{\Phi(s(b-a))} \leq \exp\left(\frac{1}{8}s^2(b-a)^2\right),$$

which proves (1.17). □

The error probability of partitioning classifier

Let $\mathcal{P}_n = \{A_{n,j},\ j = 1, 2, \dots\}$ be a cubic partition of $\mathbb{R}^d$. Put

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in A\}} Y_i$$

and

$$\nu(A) = \mathbb{E}\{\nu_n(A)\}.$$

The partitioning classification rule g_n is defined by

$$g_n(\mathbf{x}) = \text{sign } \nu_n(A) \text{ if } \mathbf{x} \in A. \tag{1.18}$$

Theorem 1.3. (Kohler and Krzyżak (2007)). *Assume that D satisfies the weak margin condition with $0 < \alpha \leq 1$ and the Lipschitz condition, and that the strong density condition is fulfilled. Then*

$$\mathbb{E}\{L(g_n)\} - L^* = O(h^{1+\alpha}) + O(1/(nh_n^d)^{(\alpha+1)/2}).$$

Proof. For the notations

$$D_n(\mathbf{x}) = \frac{\nu_n(A)}{\mu(A)} \text{ if } \mathbf{x} \in A$$

and

$$\bar{D}_n(\mathbf{x}) = \frac{\nu(A)}{\mu(A)} \text{ if } \mathbf{x} \in A,$$

we get

$$\begin{aligned}
\mathbb{E}\{L(g_n)\} - L^* &= \int \mathbb{E}\left\{\mathbb{I}_{\{sign\, D_n(\mathbf{x}) \neq sign\, D(\mathbf{x})\}}\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&\leq \int \mathbb{E}\left\{\mathbb{I}_{\{|D_n(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|\}}\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&\leq \int \mathbb{E}\left\{\mathbb{I}_{\{|D_n(\mathbf{x}) - \bar{D}_n(\mathbf{x})| \geq |D(\mathbf{x})|/2\}}\right\} |D(x)| \mu(d\mathbf{x}) \\
&+ \int \mathbb{I}_{\{|\bar{D}_n(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|/2\}} |D(\mathbf{x})| \mu(d\mathbf{x}).
\end{aligned}$$

The Lipschitz condition and the margin condition imply that

$$\begin{aligned}
&\int \mathbb{I}_{\{|\bar{D}_n(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|/2\}} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&= \int \mathbb{I}_{\{Ch_n \geq |\bar{D}_n(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|/2\}} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&\leq \int \mathbb{I}_{\{Ch_n \geq |D(\mathbf{x})|/2\}} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&= O(h_n^{1+\alpha}).
\end{aligned}$$

$D_n(\mathbf{x}) - \bar{D}_h(\mathbf{x})$ is an average of i.i.d. bounded random variables with bound

$$1/\mu(A_n(\mathbf{x}))$$

and with variance less than

$$1/\mu(A_n(\mathbf{x})).$$

Thus, from the Bernstein inequality (Lemma 1.3) we get that

$$\begin{aligned}
&\int \mathbb{E}\left\{\mathbb{I}_{\{|D_n(\mathbf{x}) - \bar{D}_n(\mathbf{x})| \geq |D(\mathbf{x})|/2\}}\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&= \int \mathbb{P}\left\{|D_n(\mathbf{x}) - \bar{D}_n(\mathbf{x})| \geq |D(\mathbf{x})|/2\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&\leq \int e^{-n|D(\mathbf{x})|^2/(8(1+|D(\mathbf{x})|)/\mu(A_n(\mathbf{x})))} |D(\mathbf{x})| \mu(d\mathbf{x}),
\end{aligned}$$

and so

$$\begin{aligned}
&\int \mathbb{E}\left\{\mathbb{I}_{\{|D_n(\mathbf{x}) - \bar{D}_n(\mathbf{x})| \geq |D(\mathbf{x})|/2\}}\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\
&\leq \int e^{-f_{min} n h_n^d |D(\mathbf{x})|^2/16} |D(\mathbf{x})| \mu(d\mathbf{x}),
\end{aligned}$$

where we applied the strong density condition. Let G be the distribution function of $|D(X)|$. Put

$$H(s) = c^* s^\alpha.$$

Because of the margin condition, we have that

$$G(s) \leq H(s).$$

Thus, by partial integration,

$$\begin{aligned}
\int e^{-f_{min}(\sqrt{nh_n^d}|D(\mathbf{x})|)^2/16} |D(\mathbf{x})| \mu(d\mathbf{x}) &= \int_0^1 e^{-f_{min}(\sqrt{nh_n^d}s)^2/16} s G(ds) \\
&\leq \int_0^1 e^{-f_{min}(\sqrt{nh_n^d}s)^2/16} s H'(s) ds \\
&= c^* \alpha \int_0^1 e^{-f_{min}(\sqrt{nh_n^d}s)^2/8} s s^{\alpha-1} ds \\
&\leq const \int_0^\infty e^{-u} u^{(\alpha-1)/2} du / (nh_n^d)^{(1+\alpha)/2} \\
&= O(1/(nh_n^d)^{(1+\alpha)/2}). \qquad (1.19)
\end{aligned}$$

□

1.6 Kernel classifier

Kernel regression estimate

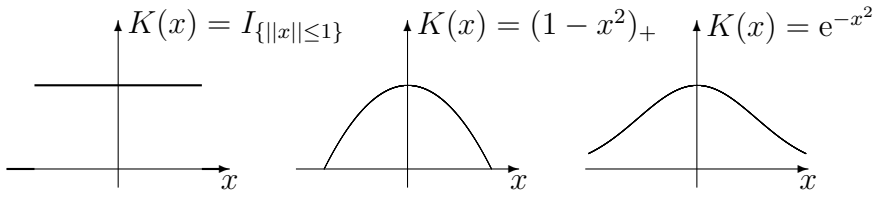

Figure 1.10: Examples for univariate kernels.

Kernel-based rules are derived from the kernel estimate in density estimation originally studied by Parzen (1962), Rosenblatt (1956), Akaike (1954), and Cacoullos (1965); and in regression estimation, introduced by Nadaraya (1964; 1970), and Watson (1964).

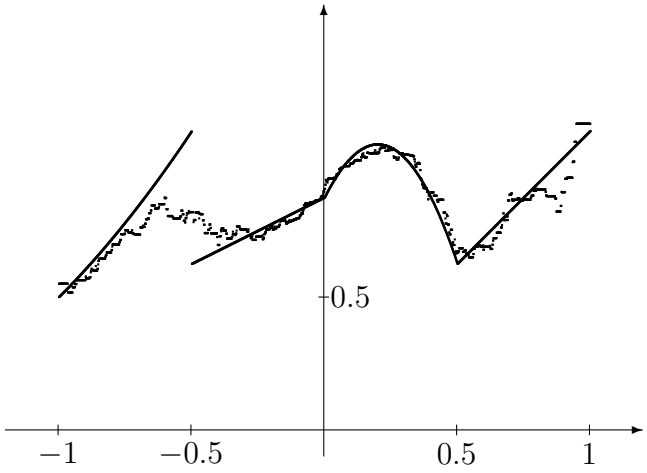

Figure 1.11: Kernel estimate for the naive kernel: $h = 0.1$, L_2 error = 0.004066.

For particular choices of K, rules of this sort have been proposed by Fix and Hodges (1951; 1952), Sebestyen (1962), Van Ryzin (1966), and Meisel (1969). Statistical analysis of these rules and/or the corresponding regression function estimate can be found in Nadaraya (1964; 1970), Rejtő and Révész (1973), Devroye and Wagner (1976; 1980a; 1980c), Greblicki (1974; 1978b; 1978a), Krzyżak and Pawlak (1984), and Devroye and

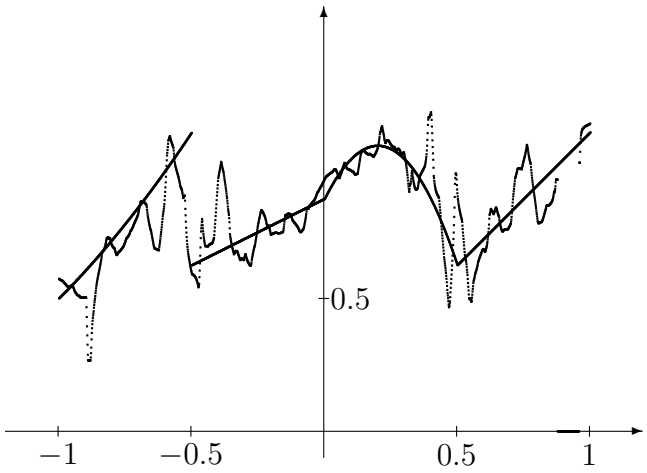

Figure 1.12: Undersmoothing for the Epanechnikov kernel: $h = 0.03$, L_2 error = 0.031560.

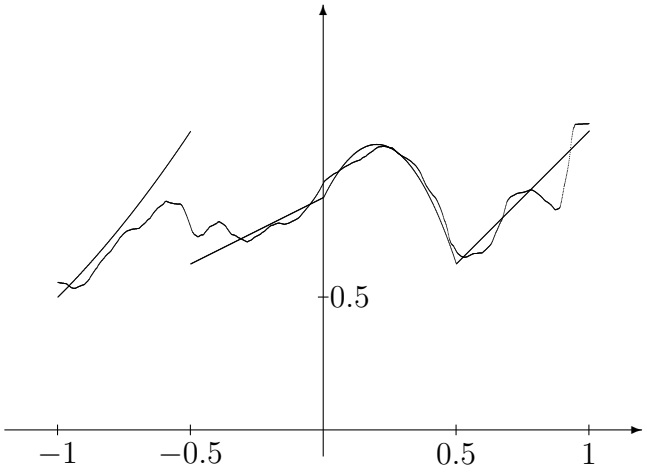

Figure 1.13: Kernel estimate for the Epanechnikov kernel: $h = 0.1$, L_2 error = 0.003608.

Krzyżak (1989). Usage of Cauchy kernels in discrimination is investigated by Arkadjew and Braverman (1966), Hand (1981), and Coomans and Broeckaert (1986).

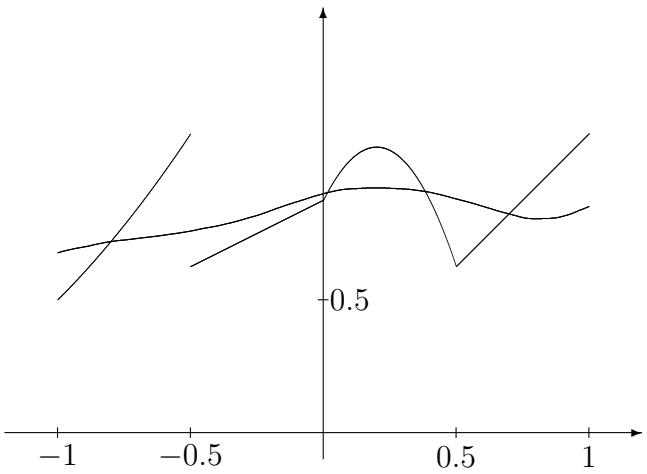

Figure 1.14: Oversmoothing for the Epanechnikov kernel: $h = 0.5$, L_2 error = 0.012551.

The kernel estimate of a regression function takes the form

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)},$$

if the denominator is nonzero, and 0 otherwise. Here the bandwidth $h_n > 0$ depends only on the sample size n, and the function $K : \mathbb{R}^d \to [0, \infty)$ is called a kernel. (See Figure 1.10 for some examples.) Usually $K(\mathbf{x})$ is "large" if $\|\mathbf{x}\|$ is "small," therefore the kernel estimate again is a local averaging estimate.

Figures 1.11–1.14 show the kernel estimate for the naive kernel

$$K(\mathbf{x}) = \mathbb{I}_{\{\|\mathbf{x}\| \leq 1\}}$$

and for the Epanechnikov kernel

$$K(\mathbf{x}) = (1 - \|\mathbf{x}\|^2)_+$$

using various choices for h_n for our simulated data.

The error probability of kernel classifier

We fix $\mathbf{x} \in \mathbb{R}^d$, and, for an $h > 0$, let the (naive) kernel estimate of $D(\mathbf{x})$ be

$$D_{n,h}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n Y_i \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h}\}} / \mu(S_{\mathbf{x},h}),$$

where $S_{\mathbf{x},h}$ denotes the sphere centered at $\mathbf{x}$ with radius h. Notice that $D_{n,h}$ is not a true estimate, because its denominator contains the unknown distribution μ. However, the corresponding plug-in classification rule defined below depends only on the sign of $D_{n,h}(\mathbf{x})$, and so μ doesn't count. The (naive) kernel classification rule is

$$g_{n,h}(\mathbf{x}) = \text{sign } D_{n,h}(\mathbf{x})$$

(cf. Devroye (1981b), Devroye and Wagner (1980b), Krzyżak (1986), Krzyżak and Pawlak (1984)).

If D is Lipschitz continuous and X is bounded then, for the L_1 error, one has that

$$\mathbb{E}\{|D(\mathbf{X}) - D_{n,h}(\mathbf{X})|\} \quad \leq \quad c_2 h + \frac{c_3}{\sqrt{nh^d}},$$

(cf. Györfi et al. (2002)), so for the choice

$$h = n^{-\frac{1}{d+2}}, \tag{1.20}$$

the L_1 upper bound implies that

$$\mathbb{E}\{L(g_{n,h})\} - L^* \leq c_4 n^{-\frac{1}{d+2}}.$$

Theorem 1.4. (Kohler and Krzyżak (2007), Döring, Györfi and Walk (2015)). *Assume that D satisfies the weak margin condition with $0 < \alpha \leq 1$ and the Lipschitz condition, and the strong density assumption is fulfilled. Then*

$$\mathbb{E}\{L(g_n)\} - L^* = O(h^{1+\alpha}) + O(1/(nh_n^d)^{(\alpha+1)/2}),$$

and so for the choice (1.20), we get that

$$\mathbb{E}\{L(g_{n,h})\} - L^* \leq c_7 n^{-\frac{1+\alpha}{d+2}}.$$

Proof. Because of (1.15), we have that the excess error probability of any plug-in classification rule has the following decomposition:

$$\begin{aligned}
\mathbb{E}\{L(g_{n,h})\} - L^* &= \mathbb{E}\left\{\int_{\{sign\, D_{n,h}(\mathbf{x}) \neq sign\, D(\mathbf{x})\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\} \\
&\leq \mathbb{E}\left\{\int_{\{|D_{n,h}(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\} \\
&\leq I_{n,h} + J_{n,h},
\end{aligned}$$

where

$$I_{n,h} = \int_{\{|\bar{D}_h(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|/2\}} |D(\mathbf{x})|\mu(d\mathbf{x})$$

and

$$J_{n,h} = \mathbb{E}\left\{\int_{\{|D_{n,h}(\mathbf{x}) - \bar{D}_h(\mathbf{x})| \geq |D(\mathbf{x})|/2\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\}$$

with $\bar{D}_h(\mathbf{x}) = \mathbb{E}\{D_{n,h}(\mathbf{x})\}$. $I_{n,h}$ is called approximation error, while $J_{n,h}$ is the estimation error. For the approximation error, notice that

$$\bar{D}_h(\mathbf{x}) = \mathbb{E}\{D_{n,h}(\mathbf{x})\} = \frac{\int_{S_{\mathbf{x},h}} D(\mathbf{z})\mu(d\mathbf{z})}{\mu(S_{\mathbf{x},h})}.$$

Therefore

$$\bar{D}_h(\mathbf{x}) - D(\mathbf{x}) = \frac{\int_{S_{\mathbf{x},h}} (D(\mathbf{z}) - D(\mathbf{x}))\mu(d\mathbf{z})}{\mu(S_{\mathbf{x},h})}.$$

By the Lipschitz condition and the margin condition

$$I_{n,h} \leq \int_{\{|D(\mathbf{x})| \leq Ch\}} |D(\mathbf{x})| \mu(d\mathbf{x}) \leq c(Ch)^{1+\alpha}.$$

Next we consider the estimation error. $D_{n,h}(\mathbf{x}) - \bar{D}_h(\mathbf{x})$ is an average of i.i.d. bounded random variables with bound

$$1/\mu(S_{\mathbf{x},h})$$

and with variance less than

$$1/\mu(S_{\mathbf{x},h}).$$

Thus, from the Bernstein inequality (Lemma 1.3) we get that

$$\begin{aligned} J_{n,h} &= \int \mathbb{E}\left\{\mathbb{I}_{\{|D_{n,h}(\mathbf{x}) - \bar{D}_n(\mathbf{x})| \geq |D(\mathbf{x})|/2\}}\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\ &= \int \mathbb{P}\left\{|D_{n,h}(\mathbf{x}) - \bar{D}_n(\mathbf{x})| \geq |D(\mathbf{x})|/2\right\} |D(\mathbf{x})| \mu(d\mathbf{x}) \\ &\leq \int e^{-n|D(\mathbf{x})|^2/(8(1+|D(\mathbf{x})|)/\mu(S_{\mathbf{x},h}))} |D(\mathbf{x})| \mu(d\mathbf{x}) \end{aligned}$$

and so

$$\begin{aligned} J_{n,h} &\leq \int e^{-f_{min} n h_n^d |D(\mathbf{x})|^2/16} |D(\mathbf{x})| \mu(d\mathbf{x}) \\ &= O(1/(nh_n^d)^{(1+\alpha)/2}), \end{aligned}$$

where we applied the strong density and the margin conditions as in (1.19). □

1.7 Nearest neighbor classifier

Nearest neighbor regression estimate

We fix $\mathbf{x} \in \mathbb{R}^d$, and reorder the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$. The reordered data sequence is denoted by

$$(\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n,n)}(\mathbf{x}), Y_{(n,n)}(\mathbf{x}))$$

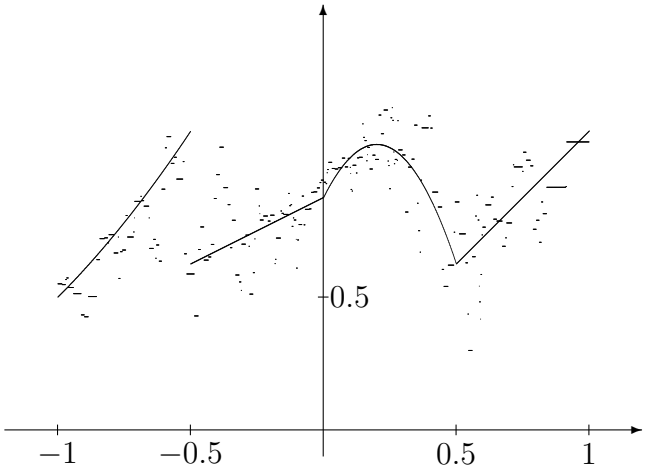

Figure 1.15: Undersmoothing: $k_n = 3$, L_2 error =0.011703.

or by

$$(\mathbf{X}_{(1,n)}, Y_{(1,n)}), \ldots, (\mathbf{X}_{(n,n)}, Y_{(n,n)})$$

if no confusion is possible. $\mathbf{X}_{(k,n)}(\mathbf{x})$ is called the kth nearest neighbor (k-NN) of $\mathbf{x}$.

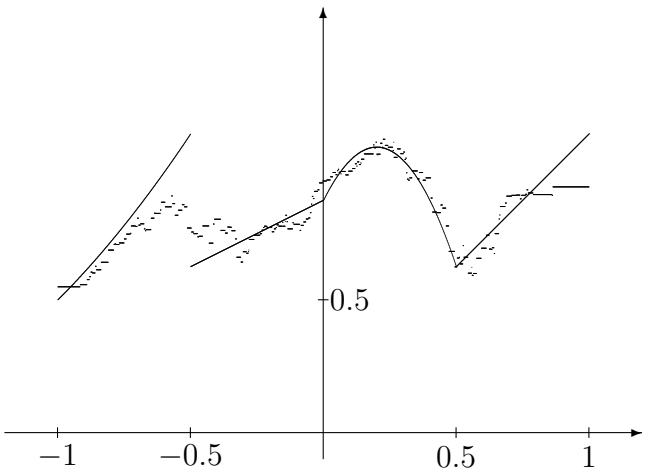

Figure 1.16: Good choice: $k_n = 12$, L_2 error =0.004247.

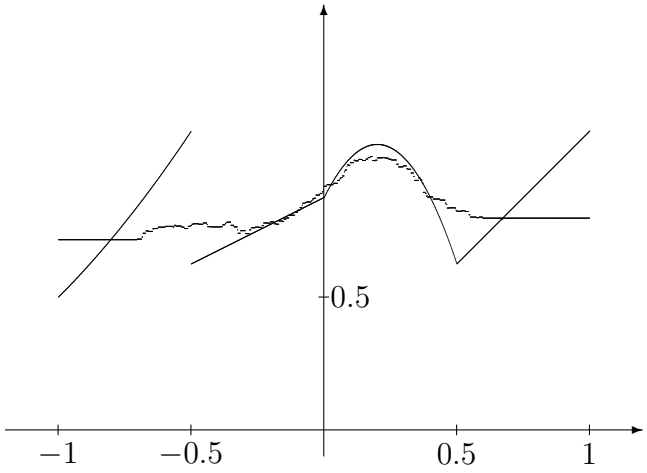

Figure 1.17: Oversmoothing: $k_n = 50$, L_2 error =0.009931.

The k_n-NN regression function estimate is defined by

$$m_n(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(\mathbf{x}).$$

If $\mathbf{X}_i$ and $\mathbf{X}_j$ are equidistant from $\mathbf{x}$, i.e., $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$, then we have a tie. There are several rules for tie breaking. For example, $\mathbf{X}_i$ might be declared "closer" if $i < j$, i.e., the tie breaking is done by indices. In the sequel we shall assume that $\mathbf{X}$ has a density, therefore, for each $\mathbf{x}$ the random variable $\|\mathbf{X} - \mathbf{x}\|^2$ is absolutely continuous, and so ties occur with probability 0.

The k-nearest neighbor rule, since its conception in 1951 and 1952 (Fix and Hodges (1951; 1952; 1991a; 1991b)), has attracted many followers and continues to be studied by many researchers. For surveys of various aspects of the nearest neighbor or related methods, see Biau and Devroye (2015), Cover and Hart (1967), Dasarathy (1991), Devijver (1980), Devroye (1981a), Devroye and Györfi (1985), Devroye and Wagner (1982), Györfi (1978) or Györfi and Györfi (1978).

Storing the n data pairs in an array and searching for the k nearest neighbors may take time proportional to nkd if done in a naive manner—the "d" accounts for the cost of one distance computation. This complexity may be reduced in terms of one or more of the three factors involved. Typically, with k and d fixed, $O(n^{1/d})$ worst-case time (Papadimitriou and Bentley (1980)) and $O(\log n)$ expected time (Friedman, Bentley, and Finkel (1977)) may be achieved. Multidimensional search trees that partition the

space and guide the search are invaluable—for this approach, see Fukunaga and Narendra (1975), Friedman, Bentley, and Finkel (1977), Niemann and Goppert (1988), Kim and Park (1986), and Broder (1990). We refer to a survey in Dasarathy (1991) for more references. Other approaches are described by Yunck (1976), Friedman, Baskett, and Shustek (1975), Vidal (1986), Sethi (1981), and Faragó, Linder, and Lugosi (1993). Generally, with preprocessing, one may considerably reduce the overall complexity in terms of n and d.

Figures 1.15 – 1.17 show k_n-NN estimates for various choices of k_n for our simulated data.

The error probability of nearest neighbor classifier

In the sequel our focus lies on the rate of convergence of the excess error probability $\mathbb{E}\{L(g_{n,k})\} - L^*$, where $g_{n,k}$ is the k-nearest neighbor rule defined as follows: Choose an integer k less than n, then the k-nearest-neighbor estimate of D is

$$D_{n,k}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} Y_{(n,i)}(\mathbf{x}), \tag{1.21}$$

and the k-nearest-neighbor classification rule is

$$g_{n,k}(\mathbf{x})) = sign\, D_{n,k}(\mathbf{x}). \tag{1.22}$$

Kohler and Krzyżak (2007) proved that under the weak margin condition, Lipschitz condition and strong density assumption we get that

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq c_5 (\log n)^{\frac{2(1+\alpha)}{d}} (k/n)^{(1+\alpha)/d} + \frac{c_6}{k^{(1+\alpha)/2}}.$$

For choice

$$k_n = c^* n^{2/(d+2)}, \tag{1.23}$$

it implies that the order of the upper bound can be smaller than:

$$(\log n)^{\frac{2(1+\alpha)}{d}} n^{-\frac{1+\alpha}{d+2}}.$$

Gadat, Klein and Mateau (2016) extended this bound such that under the weak margin condition, Lipschitz condition and the so called strong minimal mass assumption they get that

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq c_5 (k/n)^{(1+\alpha)/d} + \frac{c_6}{k^{(1+\alpha)/2}},$$

which results in the order

$$n^{-\frac{1+\alpha}{d+2}}. \tag{1.24}$$

Audibert and Tsybakov (2005) showed that (1.24) is the minimax optimal rate of convergence for the class of Lipschitz continuous D, i.e., (1.24) can be the lower bound for *any* classifier.

Hall, Park and Samworth (2008), and Samworth (2012) considered the case when the conditional densities of X given Y are twice differentiable and the density f satisfies the strong density assumption. Under some additional conditions on B_0

$$\mathbb{E}\{L(g_{n,k})\} - L^* \leq c_7(k/n)^{4/d} + \frac{c_8}{k},$$

which implies in the order

$$n^{-\frac{4}{d+4}}. \tag{1.25}$$

Audibert and Tsybakov (2005) showed that the order

$$n^{-\frac{2(1+\alpha)}{d+4}} \tag{1.26}$$

is the minimax optimal rate of convergence for the class of regression functions D, which are differentiable and the partial derivatives are Lipschitz continuous. The conditions in (2008) and (2012) imply that the strong margin condition with $\alpha = 1$, therefore (1.25) is the minimax optimal rate of convergence for this class.

Theorem 1.5. *Assume that* $\mathbf{X}$ *has a density,* D *satisfies the weak margin condition with* $0 < \alpha \leq 1$ *and the modified Lipschitz condition. Then*

$$\mathbb{E}\{L(g_{n,k})\} - L^* = O(1/k^{(1+\alpha)/2}) + O((k/n)^{(\alpha+1)/d}).$$

Proof. Because of (1.15), we have the following decomposition of the excess error probability:

$$\begin{aligned}
\mathbb{E}\{L(g_{n,k})\} - L^* &= \mathbb{E}\left\{\int_{\{sign\, D_{n,k}(\mathbf{x}) \neq sign\, D(\mathbf{x})\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\} \\
&= \mathbb{E}\left\{\int_{\{|D_{n,k}(\mathbf{x}) - D(\mathbf{x})| \geq |D(\mathbf{x})|\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\} \\
&\leq I_{n,k} + J_{n,k},
\end{aligned}$$

where

$$I_{n,k} = \mathbb{E}\left\{\int_{\{|D_{n,k}(\mathbf{x})-\bar{D}_{n,k}(\mathbf{x})|\geq|D(\mathbf{x})|/2\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\}$$

and

$$J_{n,k} = \mathbb{E}\left\{\int_{\{|\bar{D}_{n,k}(\mathbf{x})-D(\mathbf{x})|\geq|D(\mathbf{x})|/2\}} |D(\mathbf{x})|\mu(d\mathbf{x})\right\}$$

with

$$\bar{D}_{n,k}(\mathbf{x}) = \mathbb{E}\{D_{n,k}(\mathbf{x}) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\}. \tag{1.27}$$

$I_{n,k}$ is called approximation error, while $J_{n,k}$ is the estimation error. Proposition 8.1 in (2015) says the following: given $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the random pairs

$$(\mathbf{X}_{(n,1)}(\mathbf{x}), Y_{(n,1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n,k)}(\mathbf{x}), Y_{(n,k)}(\mathbf{x}))$$

are independent, and

$$\mathbb{E}\{Y_{(n,i)}(\mathbf{x}) - D(\mathbf{X}_{(n,i)}(\mathbf{x})) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\} = 0.$$

Therefore, the Hoeffding inequality (Lemma 1.4) implies that

$$\begin{aligned}
&\mathbb{P}\{|D_{n,k}(\mathbf{x}) - \bar{D}_{n,k}(\mathbf{x})| \geq |D(\mathbf{x})|/2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\} \\
&= \mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k}(Y_{(n,i)}(\mathbf{x}) - D(\mathbf{X}_{(n,i)}(\mathbf{x})))\right| \geq |D(\mathbf{x})|/2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right\} \\
&\leq 2e^{-k|D(\mathbf{x})|^2/32}.
\end{aligned}$$

Then

$$J_{n,k} \leq 2\int |D(\mathbf{x})|e^{-k|D(\mathbf{x})|^2/32}\mu(d\mathbf{x}).$$

The weak margin condition with α means that

$$G(t) := \mathbb{P}\{0 < |D(X)| \leq t\} \leq c^* \cdot t^\alpha,\ 0 \leq t \leq 1.$$

This implies that

$$\begin{aligned}
&\int |D(\mathbf{x})|\Phi\left(-\sqrt{k}|D(\mathbf{x})|/2\right)\mu(d\mathbf{x}) = \int_0^1 s\Phi\left(-\sqrt{k}s/2\right)G(ds) \\
&= s\Phi\left(-\sqrt{k}s/2\right)G(s)\Big|_0^1 - \int_0^1 \left[\Phi\left(-\sqrt{k}s/2\right) - s\frac{\sqrt{k}}{2}\Phi'\left(-\sqrt{k}s/2\right)\right]G(s)ds \\
&\leq \Phi\left(-\sqrt{k}/2\right) + \int_0^{\sqrt{k}} \frac{u}{2}\Phi'\left(-u/2\right)c^*u^\alpha du k^{-(\alpha+1)/2} = O(k^{-(\alpha+1)/2}).
\end{aligned}$$

For i.i.d. uniformly distributed $U_1, \ldots, U_n$, let $U_{(1,n)}, \ldots, U_{(n,n)}$ denote the corresponding order statistic. From Section 1.2 in Biau and Devroye (2015) we have that

$$\mu(S_{\mathbf{x},\|\mathbf{x}-\mathbf{X}_{(n,k)}(\mathbf{x})\|}) \stackrel{\mathcal{D}}{=} U_{(k,n)}. \tag{1.28}$$

Because of

$$\begin{aligned} |D(\mathbf{x}) - \bar{D}_{n,k}(\mathbf{x})| &= \left| D(\mathbf{x}) - \frac{1}{k}\sum_{i=1}^{k} D(\mathbf{X}_{(n,i)}(\mathbf{x})) \right| \\ &\leq \frac{1}{k}\sum_{i=1}^{k} |D(\mathbf{x}) - D(\mathbf{X}_{(n,i)}(\mathbf{x}))| \end{aligned}$$

the modified Lipschitz condition together with (1.28) implies that

$$\begin{aligned} &\mathbb{P}\left\{|D(\mathbf{x})|/2 < |D(\mathbf{x}) - \bar{D}_{n,k}(\mathbf{x})|\right\} \\ &\leq \mathbb{P}\left\{ |D(\mathbf{x})|/2 < C^* \frac{1}{k}\sum_{i=1}^{k} \mu(S_{\mathbf{x},\|\mathbf{x}-\mathbf{X}_{(n,i)}(\mathbf{x})\|})^{1/d} \right\} \\ &\leq \mathbb{P}\left\{ |D(\mathbf{x})|/2 < C^* \mu(S_{\mathbf{x},\|\mathbf{x}-\mathbf{X}_{(n,k)}(x)\|})^{1/d} \right\} \\ &= \mathbb{P}\left\{ |D(\mathbf{x})|/2 < C^* U_{(k,n)}^{1/d} \right\} \\ &= \mathbb{P}\left\{ |D(\mathbf{x})|^d/(2C^*)^d < U_{(k,n)} \right\}. \end{aligned} \tag{1.29}$$

Without loss of generality, assume that $C^* \geq 1/2$. Then

$$\begin{aligned} &\mathbb{P}\left\{|D(\mathbf{x})|/2 < |D(\mathbf{x}) - \bar{D}_{n,k}(\mathbf{x})|\right\} \\ &\leq \mathbb{P}\left\{ \sum_{i=1}^{n} \mathbb{I}_{\{U_i \leq |D(\mathbf{x})|^d/(2C^*)^d\}} < k \right\} \\ &\leq \mathbb{I}_{\{|D(\mathbf{x})|^d/(2C^*)^d \geq 2k/n\}} \mathbb{P}\left\{ \sum_{i=1}^{n} \mathbb{I}_{\{U_i \leq |D(\mathbf{x})|^d/(2C^*)^d\}} < \frac{n}{2}|D(\mathbf{x})|^d/(2C^*)^d \right\} \\ &+ \mathbb{I}_{\{|D(\mathbf{x})|^d/(2C^*)^d < 2k/n\}} \\ &\leq \mathbb{I}_{\{|D(\mathbf{x})|^d/(2C^*)^d \geq 2k/n\}} e^{-\frac{1-\log 2}{2} n|D(\mathbf{x})|^d/(2C^*)^d} + \mathbb{I}_{\{|D(\mathbf{x})|^d/(2C^*)^d < 2k/n\}} \\ &\leq e^{-(1-\log 2)k} + \mathbb{I}_{\{|D(\mathbf{x})|^d/(2C^*)^d < 2k/n\}}, \end{aligned} \tag{1.30}$$

where the third inequality follows from Chernoff's exponential inequality (Lemma 1.2). Applying the weak margin condition, we get

$$\begin{aligned} I_{n,k} &= \int |D(\mathbf{x})| \mathbb{P}\left\{|D(\mathbf{x})|/2 < |D(\mathbf{x}) - \bar{D}_{n,k}(\mathbf{x})|\right\} \mu(d\mathbf{x}) \\ &\le e^{-(1-\log 2)k} + O((k/n)^{(\alpha+1)/d}). \end{aligned} \tag{1.31}$$

□

1.8 Empirical error minimization

Selection of classifier

In this section we consider the selection of a classification function from a class $\mathcal{G}$ of functions. If a class $\mathcal{G}$ of classifiers is given, then it is tempting to pick the one that minimizes an estimate of the error probability over the class. A good method should pick a classifier with an error probability that is close to the minimal error probability in the class. Here we require much more than distribution-free performance bounds of the error estimator for each of the classifiers in the class. Intuitively, if we can estimate the error probability for the classifiers in $\mathcal{G}$ *uniformly* well, then the classification function that minimizes the estimated error probability is likely to have an error probability that is close to the best in the class. To certify this intuition, consider the following situation: Let $\mathcal{G}$ be a class of classifiers, that is, a class of mappings of the form $g : \mathbb{R}^d \to \{-1, 1\}$. Assume that the empirical error

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{j=1}^{n} I_{\{g(\mathbf{X}_j) \neq Y_j\}}$$

is used to estimate the error probability

$$L(g) = \mathbb{P}\{g(\mathbf{X}) \neq Y\}$$

of each classifier $g \in \mathcal{G}$. Denote by g_n the classifier that minimizes the empirical error over the class:

$$\widehat{L}_n(g_n) \le \widehat{L}_n(g) \qquad \text{for all } g \in \mathcal{G}.$$

Thus, g_n is the classifier that, according to the data $\mathcal{D}_n$, "looks best" among the classifiers in $\mathcal{G}$. This idea of minimizing the empirical risk in the construction of a rule was

developed to great extent by Vapnik and Chervonenkis (1971; 1974c; 1974a; 1974b). In practice, finding an empirically optimal classifier is often computationally very expensive.

Intuitively, the selected classifier g_n should be good in the sense that its *true* error probability $L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y \mid \mathcal{D}_n\}$ is expected to be close to the optimal error probability within the class.

For the error probability

$$L(g_n) = \mathbb{P}\left\{g_n(\mathbf{X}) \neq Y \mid \mathcal{D}_n\right\}$$

of the selected rule we have:

Lemma 1.5. (Vapnik and Chervonenkis (1974c); see also Devroye (1988)).

$$L(g_n) - \inf_{g\in\mathcal{G}} L(g) \leq 2\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)|,$$

$$|\widehat{L}_n(g_n) - L(g_n)| \leq \sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)|.$$

Proof.

$$\begin{aligned} L(g_n) - \inf_{g\in\mathcal{G}} L(g) &= L(g_n) - \widehat{L}_n(g_n) + \widehat{L}_n(g_n) - \inf_{g\in\mathcal{G}} L(g) \\ &\leq L(g_n) - \widehat{L}_n(g_n) + \sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)| \\ &\leq 2\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)|. \end{aligned}$$

The second inequality is trivially true. □

We see that upper bounds for $\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)|$ provide us with upper bounds for two things simultaneously:

(1) An upper bound for the suboptimality of g_n within $\mathcal{G}$, that is, a bound for $L(g_n) - \inf_{g\in\mathcal{G}} L(g)$.

(2) An upper bound for the error $|\widehat{L}_n(g_n) - L(g_n)|$ committed when $\widehat{L}_n(g_n)$ is used to estimate the probability of error $L(g_n)$ of the selected rule.

In other words, by bounding $\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)|$, we kill two flies at once. It is particularly useful to know that even though $\widehat{L}_n(g_n)$ is usually optimistically biased, it is within given bounds of the unknown probability of error with g_n, and that no other test

sample is needed to estimate this probability of error. Whenever our bounds indicate that we are close to the optimum in $\mathcal{G}$, we must at the same time have a good estimate of the probability of error, and vice versa.

The difference

$$L(g_n) - \inf_{g \in \mathcal{G}} L(g)$$

is the quantity that primarily interests us in the sequel. This difference may be bounded in a distribution-free manner, and a rate of convergence results that only depends on the structure of $\mathcal{G}$. While this is very exciting, we must add that $L(g_n)$ may be far away from the Bayes error L^*. Note that

$$L(g_n) - L^* = \left(L(g_n) - \inf_{g \in \mathcal{G}} L(g) \right) + \left(\inf_{g \in \mathcal{G}} L(g) - L^* \right).$$

The size of $\mathcal{G}$ is a compromise: when $\mathcal{G}$ is large, $\inf_{g \in \mathcal{G}} L(g)$ may be close to L^*, but the *estimation error*

$$L(g_n) - \inf_{g \in \mathcal{G}} L(g)$$

is probably large as well. If $\mathcal{G}$ is too small, there is no hope to make the *approximation error*

$$\inf_{g \in \mathcal{G}} L(g) - L^*$$

small. For example, if $\mathcal{G}$ is the class of all decision functions, then we can always find a classifier in $\mathcal{G}$ with zero empirical error, but it may have arbitrary values outside of the points $X_1, \ldots, X_n$. For example, an empirically optimal classifier is

$$g_n(x) = \begin{cases} Y_i & \text{if } \mathbf{x} = \mathbf{X}_i,\ i = 1, \ldots, n \\ -1 & \text{otherwise.} \end{cases}$$

This is clearly not what we are looking for. This phenomenon is called *overfitting*, as the overly large class $\mathcal{G}$ overfits the data. We will give precise conditions on $\mathcal{G}$ that allow us to avoid this anomaly. The choice of $\mathcal{G}$ such that $\inf_{g \in \mathcal{G}} L(g)$ is close to L^* has been the subject of various chapters on consistency—just assume that $\mathcal{G}$ is allowed to grow with n in some manner.

Finite class $\mathcal{G}$

Here we take the point of view that $\mathcal{G}$ is fixed, and that we have to live with the functions in $\mathcal{G}$. The best we may then hope for is to minimize $L(g_n) - \inf_{g \in \mathcal{G}} L(g)$.

As a simple, but interesting application of Lemma 1.5 we consider the case when the class $\mathcal{G}$ contains finitely many classifiers.

Theorem 1.6. *Assume that the cardinality of $\mathcal{G}$ is bounded by N. Then we have for all $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)| > \epsilon\right\} \leq 2Ne^{-2n\epsilon^2}.$$

PROOF.

$$\begin{aligned}\mathbb{P}\left\{\sup_{g\in\mathcal{G}} |\widehat{L}_n(g) - L(g)| > \epsilon\right\} &\leq \sum_{g\in\mathcal{G}} \mathbb{P}\left\{|\widehat{L}_n(g) - L(g)| > \epsilon\right\} \\ &\leq 2Ne^{-2n\epsilon^2},\end{aligned}$$

where we used Hoeffding's inequality, and the fact that the random variable $n\widehat{L}_n(g)$ is binomially distributed with parameters n and $L(g)$. □

Consider a finite collection $\mathcal{G}$, and assume that one of the classifiers in $\mathcal{G}$ has zero error probability, that is, $\min_{g\in\mathcal{G}} L(g) = 0$. Then clearly, $\widehat{L}_n(g_n) = 0$ with probability one. We then have the following performance bound:

Theorem 1.7. (VAPNIK AND CHERVONENKIS (1974C)). *Assume that the cardinality $|\mathcal{G}|$ of $\mathcal{G}$ is finite, and $\min_{g\in\mathcal{G}} L(g) = 0$. Then for every n and $\epsilon > 0$,*

$$\mathbb{P}\{L(g_n) > \epsilon\} \leq |\mathcal{G}|e^{-n\epsilon},$$

and

$$\mathbb{E}\{L(g_n)\} \leq \frac{1 + \log|\mathcal{G}|}{n}.$$

PROOF. Clearly,

$$\begin{aligned}\mathbb{P}\{L(g_n) > \epsilon\} &\leq \mathbb{P}\left\{\max_{g\in\mathcal{G}:\widehat{L}_n(g)=0} L(g) > \epsilon\right\} \\ &= \mathbb{E}\left\{I_{\left\{\max_{g\in\mathcal{G}:\widehat{L}_n(g)=0} L(g)>\epsilon\right\}}\right\} \\ &= \mathbb{E}\left\{\max_{g\in\mathcal{G}} I_{\{\widehat{L}_n(g)=0\}} I_{\{L(g)>\epsilon\}}\right\} \\ &\leq \sum_{g\in\mathcal{G}:L(g)>\epsilon} \mathbb{P}\left\{\widehat{L}_n(g) = 0\right\} \\ &\leq |\mathcal{G}|(1-\epsilon)^n,\end{aligned}$$

since the probability that no $(\mathbf{X}_i, Y_i)$ pair falls in the set $\{(\mathbf{x}, y) : g(\mathbf{x}) \neq y\}$ is less than $(1-\epsilon)^n$ if the probability of the set is larger than ϵ. The probability inequality of the theorem follows from the simple inequality $1 - x \leq e^{-x}$.

To bound the expected error probability, note that for any $u > 0$,

$$\begin{aligned}
\mathbb{E}\{L(g_n)\} &= \int_0^\infty \mathbb{P}\{L(g_n) > t\} dt \\
&\leq u + \int_u^\infty \mathbb{P}\{L(g_n) > t\} dt \\
&\leq u + |\mathcal{G}| \int_u^\infty e^{-nt} dt \\
&= u + \frac{|\mathcal{G}|}{n} e^{-nu}.
\end{aligned}$$

Since u was arbitrary, we may choose it to minimize the obtained upper bound. The optimal choice is $u = \log |\mathcal{G}|/n$, which yields the desired inequality. □

Theorem 1.7 shows that empirical selection works very well if the sample size n is much larger than the logarithm of the size of the family $\mathcal{G}$. Unfortunately, the assumption on the distribution of $(\mathbf{X}, Y)$, that is, that $\min_{g \in \mathcal{G}} L(g) = 0$, is very restrictive. In the sequel we drop this assumption, and deal with the distribution-free problem.

One of our main tools is taken from Lemma 1.5:

$$L(g_n) - \inf_{g \in \mathcal{G}} L(g) \leq 2 \sup_{g \in \mathcal{G}} \left| \widehat{L}_n(g) - L(g) \right|.$$

This leads to the study of *uniform* deviations of relative frequencies from their probabilities by the following simple observation: let ν be a probability measure of $(\mathbf{X}, Y)$ on $\mathbb{R}^d \times \{-1, 1\}$, and let ν_n be the empirical measure based upon D_n. That is, for any fixed measurable set $A \subset \mathbb{R}^d \times \{-1, 1\}$,

$$\nu(A) = \mathbb{P}\{(\mathbf{X}, Y) \in A\},$$

and

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^n I_{\{(\mathbf{X}_i, Y_i) \in A\}}.$$

Then

$$L(g) = \nu(\{(\mathbf{x}, y) : g(\mathbf{x}) \neq y\})$$

is just the ν-measure of the set of pairs $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$, where $g(\mathbf{x}) \neq y$. Formally, $L(g)$ is the ν-measure of the set

$$\{\{\mathbf{x} : g(\mathbf{x}) = 1\} \times \{-1\}\} \bigcup \{\{\mathbf{x} : g(\mathbf{x}) = -1\} \times \{1\}\}.$$

Similarly,

$$\widehat{L}_n(g) = \nu_n(\{(\mathbf{x}, y) : g(\mathbf{x}) \neq y\}).$$

Thus,

$$\sup_{g \in \mathcal{G}} |\widehat{L}_n(g) - L(g)| = \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|,$$

where $\mathcal{A}$ is the collection of all sets

$$\{\{\mathbf{x} : g(\mathbf{x}) = 1\} \times \{-1\}\} \bigcup \{\{\mathbf{x} : g(\mathbf{x}) = -1\} \times \{1\}\}, \quad g \in \mathcal{G}.$$

For a fixed set A, for any probability measure ν, by the law of large numbers

$$\nu_n(A) - \nu(A) \to 0$$

almost surely as $n \to \infty$. Moreover, by Hoeffding's inequality,

$$\mathbb{P}\{|\nu_n(A) - \nu(A)| > \epsilon\} \leq 2e^{-2n\epsilon^2}.$$

However, it is a much harder problem to obtain such results for $\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|$. If the class of sets $\mathcal{A}$ (or, analogously, in the pattern recognition context, $\mathcal{G}$) is of finite cardinality, then the union bound trivially gives

$$\mathbb{P}\left\{\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\} \leq 2|\mathcal{A}|e^{-2n\epsilon^2}.$$

Infinite class $\mathcal{G}$

If the class $\mathcal{G}$ is not finite, then we need *uniform law of large numbers*, which is part of advanced probability theory. Therefore, in the sequel we omit the proofs. If $\mathcal{A}$ contains infinitely many sets (as in many of the interesting cases) then the problem becomes non-trivial, spawning a vast literature. The most powerful weapons to attack these problems are distribution-free large deviation-type inequalities first proved by Vapnik and Chervonenkis (1971) in their pioneering work. However, in some situations, we can handle the problem in a much simpler way.

Definition 1.1. *Let $\mathcal{A}$ be a collection of measurable sets. For $(\mathbf{z}_1, \ldots, \mathbf{z}_n) \in \{\mathbb{R}^d\}^n$, let $N_{\mathcal{A}}(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be the number of different sets in*

$$\{\{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \cap A; A \in \mathcal{A}\}.$$

The n-th shatter coefficient of $\mathcal{A}$ is

$$s(\mathcal{A}, n) = \max_{(\mathbf{z}_1, \ldots, \mathbf{z}_n) \in \{\mathbb{R}^d\}^n} N_{\mathcal{A}}(\mathbf{z}_1, \ldots, \mathbf{z}_n).$$

That is, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$.

The shatter coefficients measure the richness of the class $\mathcal{A}$. Clearly, $s(\mathcal{A}, n) \leq 2^n$, as there are 2^n subsets of a set with n elements. If $N_{\mathcal{A}}(\mathbf{z}_1, \ldots, \mathbf{z}_n) = 2^n$ for some $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, then we say that $\mathcal{A}$ *shatters* $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. If $s(\mathcal{A}, n) < 2^n$, then any set of n points has a subset such that there is no set in $\mathcal{A}$ that contains exactly that subset of the n points. Clearly, if $s(\mathcal{A}, k) < 2^k$ for some integer k, then $s(\mathcal{A}, n) < 2^n$ for all $n > k$. The first time when this happens is important:

Definition 1.2. *Let $\mathcal{A}$ be a collection of sets with $|\mathcal{A}| \geq 2$. The largest integer $k \geq 1$ for which $s(\mathcal{A}, k) = 2^k$ is denoted by $V_{\mathcal{A}}$, and it is called the Vapnik-Chervonenkis dimension (or* VC *dimension) of the class $\mathcal{A}$. If $s(\mathcal{A}, n) = 2^n$ for all n, then by definition, $V_{\mathcal{A}} = \infty$.*

Chapter 13 in Devroye, Györfi, Lugosi (1996) contains some examples on VC dimensions. In the most interesting cases, the class $\mathcal{G}$ of decision functions g is derived from a finite-dimensional vector space $\mathcal{F}$ of real functions $f : \mathbb{R}^d \to \mathbb{R}$ such that

$$g(\mathbf{x}) := sign\,(f(\mathbf{x})).$$

Theorem 1.8. (Steele (1975), Dudley (1978)). *Let $\mathcal{F}$ be a finite-dimensional vector space of real functions on $\mathbb{R}^d$. The class of sets*

$$\mathcal{A} = \{\{\mathbf{x} : f(\mathbf{x}) \geq 0\} : f \in \mathcal{F}\}$$

has VC *dimension $V_{\mathcal{A}} \leq r$, where $r = dimension(\mathcal{F})$.*

The distribution-free performance bound for finite VC dimension is formulated as follows:

Theorem 1.9. (Vapnik and Chervonenkis (1971)). *If the class $\mathcal{G}$ has finite* VC *dimension $V_{\mathcal{G}} > 2$, then*

$$\mathbb{E}\{L(g_n)\} - \inf_{g \in \mathcal{G}} L(g) \leq 16\sqrt{\frac{V_{\mathcal{G}} \log n + 4}{2n}}.$$

Linear and generalized linear discrimination

For linear discrimination, we split the space by a hyperplane and assign a different class to each halfspace. Such rules offer tremendous advantages—they are easy to interpret as each decision is based upon the sign of

$$\sum_{i=1}^{d} a_i x^{(i)} + a_0,$$

where $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$ and the a_i's are weights. Thus, the corresponding linear discriminant function g is defined by

$$g(\mathbf{x}) = sign\left(\sum_{i=1}^{d} a_i x^{(i)} + a_0\right).$$

The weight vector determines the relative importance of the components. The decision is also easily implemented—in a standard software solution, the time of a decision is proportional to d—and the prospect that a small chip can be built to make a virtually instantaneous decision is particularly exciting.

Rosenblatt (1962) realized the tremendous potential of such linear rules and called them *perceptrons*. Changing one or more weights as new data arrive allows us to quickly and easily adapt the weights to new situations. Training or learning patterned after the human brain thus became a reality.

Theorem 1.10. *Assume that* $\mathbf{X}$ *has a density. Let* $\mathcal{G}$ *be the class of linear discriminant functions. If* g_n *is found by empirical error minimization and* $n \geq d$*, then*

$$\mathbb{E}\left\{L(g_n)\right\} - \inf_{g \in \mathcal{G}} L(g) \leq \sqrt{\frac{2((d+1)\log n + (2d+2))}{n}}.$$

Theorem 1.10 is slightly better than the combination of Theorems 1.8 and 1.9.

The extension of linear classifier, called generalized linear classifier, have their roots in the Fourier series estimate or other series estimates of an unknown density, potential function methods. All these estimators can be put into the following form: classify $\mathbf{x}$ as

$$g(\mathbf{x}) = sign\left(\sum_{j=1}^{k} a_{n,j}\psi_j(\mathbf{x})\right),$$

where the ψ_j's are fixed functions, forming a base for the series estimate, $a_{n,j}$ is a fixed function of the training data, and k controls the amount of smoothing. When the ψ_j's are

the usual trigonometric basis, then this leads to the Fourier series classifier studied by Greblicki and Pawlak (1981; 1982). When the ψ_j's form an orthonormal system based upon Hermite polynomials, we obtain the classifiers studied by Greblicki (1981), and Greblicki and Pawlak (1983; 1985). When $\{\psi_j(\mathbf{x})\}$ is the collection of all products of components of $\mathbf{x}$ (such as $1, (x^{(i)})^k, (x^{(i)})^k (x^{(j)})^l$, etcetera), we obtain the polynomial method of Specht (1971).

Further properties of linear and generalized linear discrimination can be found in Chapters 4 and 17 of Devroye, Györfi, Lugosi (1996).

Tree classifier and data dependent partitioning

Classification trees partition $\mathbb{R}^d$ into regions, often hyperrectangles parallel to the axes. Among these, the most important are the binary classification trees, since they have just two children per node and are thus easiest to manipulate and update. We recall the simple terminology of books on data structures. The top of a binary tree is called the *root*. Each *node* has either no child (in that case it is called a *terminal node* or *leaf*), a *left child*, a *right child*, or a left child and a right child. Each node is the root of a tree itself. The trees rooted at the children of a node are called the left and right *subtrees* of that node. The *depth* of a node is the length of the *path* from the node to the root. The *height* of a tree is the maximal depth of any node.

Trees with more than two children per node can be reduced to binary trees by a simple device—just associate a left child with each node by selecting the oldest child in the list of children. Call the right child of a node its next sibling. The new binary tree is called the oldest-child/next-sibling binary tree (see, e.g., Cormen, Leiserson, and Rivest (1990) for a general introduction). We only mention this particular mapping because it enables us to only consider binary trees for simplicity.

In a *classification tree*, each node represents a set in the space $\mathbb{R}^d$. Also, each node has exactly two or zero children. If a node u represents the set A and its children u', u'' represent A' and A'', then we require that $A = A' \cup A''$ and $A' \cap A'' = \emptyset$. The root represents $\mathbb{R}^d$, and the leaves, taken together, form a *partition* of $\mathbb{R}^d$.

Assume that we know $\mathbf{x} \in A$. Then the question "is $\mathbf{x} \in A$?" should be answered in a computationally simple manner so as to conserve time. Therefore, if $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$, we may just limit ourselves to questions of the following forms:

(i) Is $x^{(i)} \leq \alpha$? This leads to *ordinary binary classification trees* with partitions into hyperrectangles.

(ii) Is $a_1 x^{(1)} + \cdots + a_d x^{(d)} \leq \alpha$? This leads to BSP *trees* (*binary space partition trees*).

Each decision is more time consuming, but the space is more flexibly cut up into convex polyhedral cells.

(iii) Is $\|\mathbf{x}-\mathbf{z}\| \leq \alpha$? (Here $\mathbf{z}$ is a point of $\mathbb{R}^d$, to be picked for each node.) This induces a partition into pieces of spheres. Such trees are called *sphere trees*.

(iv) Is $\psi(\mathbf{x}) \geq 0$? Here, ψ is a nonlinear function, different for each node. Every classifier can be thought of as being described in this format—decide class one if $\psi(\mathbf{x}) \geq 0$. However, this misses the point, as tree classifiers should really be built up from fundamental atomic operations and queries such as those listed in (i)–(iii). We will not consider such trees any further.

We associate a class in some manner with each leaf in a classification tree. The tree structure is usually data dependent, as well, and indeed, it is in the construction itself where methods differ. If a leaf represents region A, then we say that the classifier g is *natural* if

$$g(x) = sign\left(\sum_{i:X_i \in A} Y_i\right) \text{ if } \mathbf{x} \in A.$$

That is, in every leaf region, we take a majority vote over all $(\mathbf{X}_i, Y_i)$'s with $\mathbf{X}_i$ in the same region. In this set-up, natural tree classifiers are but special cases of data-dependent partitioning rules.

Regular histograms can also be thought of as natural binary tree classifiers—the construction and relationship is obvious. However, as $n \to \infty$, histograms change size, and usually, histogram partitions are not nested as n grows. Trees offer the exciting perspective of fully dynamic classification—as data are added, we may update the tree slightly, say, by splitting a leaf or so, to obtain an updated classifier.

The most compelling reason for using binary tree classifiers is to explain complicated data and to have a classifier that is easy to analyze and understand. In fact, expert system design is based nearly exclusively upon decisions obtained by going down a binary classification tree. Some argue that binary classification trees are preferable over BSP trees for this simple reason. As argued in Breiman, Friedman, Olshen, and Stone (1984), trees allow mixing component variables that are heterogeneous—some components may be of a nonnumerical nature, others may represent integers, and still others may be real numbers.

In 1984, Breiman, Friedman, Olshen, and Stone presented their CART program for constructing classification trees with perpendicular splits. One of the key ideas in their approach is the notion that trees should be constructed from the bottom up, by combining

small subtrees. The starting point is a tree with $n+1$ leaf regions defined by a partition of the space based on the n data points. Such a tree is much too large and is pruned by some methods that will not be explored here. When constructing a starting tree, a certain splitting criterion is applied recursively. The criterion determines which rectangle should be split, and where the cut should be made. To keep the classifier invariant under monotone transformation of the coordinate axes, the criterion should only depend on the coordinatewise ranks of the points, and their labels. Typically the criterion is a function of the numbers of points labeled by -1 and 1 in the rectangles after the cut is made.

There are many examples for data dependent partitioning, for example, statistically equivalent blocks, partitioning rules based on clustering, data-based scaling, classification trees, etc., see Chapters 20 and 21 in Devroye, Györfi, Lugosi (1996).

Neural network

The linear discriminant or perceptron makes a decision

$$\phi(\mathbf{x}) = sign\ (\psi(\mathbf{x}))$$

based upon a linear combination $\psi(\mathbf{x})$ of the inputs,

$$\psi(\mathbf{x}) = c_0 + \sum_{i=1}^{d} c_i x^{(i)} = c_0 + \mathbf{c}^T\mathbf{x}, \tag{1.32}$$

where the c_i's are weights, $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})^T$, and $\mathbf{c} = (c_1, \ldots, c_d)^T$. This is called a neural network without hidden layers.

In a (feed-forward) neural network with one hidden layer, one takes

$$\psi(\mathbf{x}) = c_0 + \sum_{i=1}^{k} c_i \sigma(\psi_i(\mathbf{x})), \tag{1.33}$$

where the c_i's are as before, and each ψ_i is of the form given in (1.32):

$$\psi_i(\mathbf{x}) = b_i + \sum_{j=1}^{d} a_{ij} x^{(j)}$$

for some constants b_i and a_{ij}. The function σ is called a *sigmoid*. We define sigmoids to be nondecreasing functions with $\sigma(x) \to -1$ as $x \downarrow -\infty$ and $\sigma(x) \to 1$ as $x \uparrow \infty$. Examples include:

(1) the *threshold* sigmoid

$$\sigma(x) = sign\,(x);$$

(2) the *standard*, or *logistic*, sigmoid

$$\sigma(x) = \frac{1 - e^{-x}}{1 + e^{-x}};$$

(3) the *arctan* sigmoid

$$\sigma(x) = \frac{2}{\pi}\arctan(x);$$

(4) the *gaussian* sigmoid

$$\sigma(x) = 2\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du - 1.$$

For early discussion of multilayer perceptrons, see Rosenblatt (1962), Barron (1975), Nilsson (1965), and Minsky and Papert (1969). Surveys may be found in Barron and Barron (1988), Ripley (1993; 1994), Hertz, Krogh, and Palmer (1991), and Weiss and Kulikowski (1991).

In the perceptron with one hidden layer, we say that there are k hidden neurons—the output of the i-th hidden neuron is $u_i = \sigma(\psi_i(\mathbf{x}))$. Thus, (1.33) may be rewritten as

$$\psi(\mathbf{x}) = c_0 + \sum_{i=1}^{k} c_i u_i,$$

which is similar in form to (1.32). We may continue this process and create multilayer feed-forward neural networks. For example, a two-hidden-layer perceptron uses

$$\psi(\mathbf{x}) = c_0 + \sum_{i=1}^{l} c_i z_i,$$

where

$$z_i = \sigma\left(d_{i0} + \sum_{j=1}^{k} d_{ij} u_j\right), \quad 1 \le i \le l,$$

and

$$u_j = \sigma\left(b_j + \sum_{i=1}^{d} a_{ji} x^{(i)}\right), \quad 1 \le j \le k,$$

and the d_{ij}'s, b_j's, and a_{ji}'s are constants. The first hidden layer has k hidden neurons, while the second hidden layer has l hidden neurons.

The step from perceptron to a one-hidden-layer neural network is nontrivial. We know that linear discriminants cannot possibly lead to universally consistent rules. Fortunately, one-hidden-layer neural networks yield universally consistent discriminants provided that we allow k, the number of hidden neurons, to grow unboundedly with n. The interest in neural networks is undoubtedly due to the possibility of implementing them directly via processors and circuits. As the hardware is fixed beforehand, one does not have the luxury to let k become a function of n, and thus, the claimed universal consistency is a moot point. We will deal with both fixed architectures and variable-sized neural networks. Because of the universal consistency of one-hidden-layer neural networks, there is little theoretical gain in considering neural networks with more than one hidden layer. There may, however, be an information-theoretic gain as the number of hidden neurons needed to achieve the same performance may be substantially reduced. In fact, we will make a case for two hidden layers, and show that after two hidden layers, little is gained for classification.

For theoretical analysis, the neural networks are rooted in a classical theorem by Kolmogorov (1957) and Lorentz (1976) which states that every continuous function f on $[0,1]^d$ can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{2d+1} F_i \left(\sum_{j=1}^{d} G_{ij}(x^{(j)}) \right),$$

where the G_{ij}'s and the F_i's are continuous functions whose form depends on f. One can show that neural networks approximate any measurable function with arbitrary precision, despite the fact that the form of the sigmoids is fixed beforehand.

Further properties of neural network can be found in Chapter 30 in Devroye, Györfi, Lugosi (1996).

Support vector machine

We can start with the setup of generalized discrimination, where $\psi_j, j = 1, 2, \ldots$ are linearly independent functions defined on $\mathbb{R}^d$, and $\mathcal{G}_k$ is the class of decision functions g of form

$$g(\mathbf{x}) = sign \left(\sum_{j=1}^{k} c_j \psi_j(\mathbf{x}) \right)$$

with weight vector $\mathbf{c} = (c_1, \ldots, c_k)$. In this way $L(g_n)$ can approach $\inf_{g \in \mathcal{G}_k} L(g)$ if n is large enough. In order to achieve consistency, one has to increase k. Furthermore,

even for fixed k, we can decrease $\inf_{g\in\mathcal{G}_k} L(g)$ by appropriate choice of the functions $\psi_j, j = 1, 2, \dots$, which is interpreted as nonlinear transformation of the feature vector $\mathbf{X}$:

$$(\psi_1(\mathbf{X}), \dots, \psi_k(\mathbf{X})).$$

Put

$$\mathcal{F} = \{f : f = \sum_{j=1}^{k} c_j \psi_j(\mathbf{x}),\ \mathbf{c} \in \mathbb{R}^k,\ k = 1, 2, \dots\}$$

and

$$\mathcal{G} = \{g : g(\mathbf{x}) = sign\, f(\mathbf{x})\}.$$

The main aim of support vector machine algorithms is to approximate $\inf_{g\in\mathcal{G}} L(g)$.

The previously mentioned empirical error minimization principle cannot be applied, because the class $\mathcal{F}$ is too large, its VC dimension is not finite. Assume that, for each $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, the function

$$K(\mathbf{x}, \mathbf{z}) := \sum_{j=1}^{\infty} \psi_j(\mathbf{x})\psi_j(\mathbf{z})$$

is well defined and finite. The function K is called kernel function. The kernel function uniquely generates the so called *Reproducing Kernel Hilbert Space* (RKHS) with the norm $\|\cdot\|$.

Furthermore, the $0-1$ loss is replaced by a continuous approximate loss ℓ. The empirical error minimization would result in a complete overfitting, therefore it is replaced by *complexity regularization*:

$$f_n = \arg\min_{f\in\mathcal{F}} \left(\frac{1}{n} \sum_{j=1}^{n} \ell(f(\mathbf{X}_j), Y_j) + \lambda_n \|f\| \right)$$

with $\lambda_n \downarrow 0$, and the corresponding decision function is defined by

$$g_n(\mathbf{x}) = sign\, f_n(\mathbf{x}).$$

Concerning the detailed theory of support vector machines, we suggest to visit the books Schölkopf and Smola (2002), Shawe-Taylor and Cristianini (2004), Steinwart and Christmann (2008), Suykens et al. (2002), Vapnik (1995) and Vapnik and Kotz (2006).

Chapter 2

Testing Simple Hypotheses

2.1 α-level tests

In this section we consider decision problems, where the consequences of the various errors are very much different. For example, if in a diagnostic problem $Y = 0$ means that the patient is OK, while $Y = 1$ means that the patient is ill, then for $Y = 0$ the false decision is that the patient is ill, which implies some superfluous medical treatment, while for $Y = 1$ the false decision means that the illness is not detected, and the patient's state may become worse. A similar situation happens for radar detection.

The event $Y = 0$ is called null hypothesis and is denoted by $\mathcal{H}_0$, and the event $Y = 1$ is called alternative hypothesis and is denoted by $\mathcal{H}_1$. The decision, the test is formulated by a set $A \subset \mathbb{R}^d$, called acceptance region such that accept $\mathcal{H}_0$ if $\mathbf{X} \in A$, otherwise reject $\mathcal{H}_0$, i.e., accept $\mathcal{H}_1$. The set A^c is called critical region.

Let P_0 and P_1 be the probability distributions of $\mathbf{X}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. There are two types of errors:

- Error of the first kind, if under the null hypothesis $\mathcal{H}_0$ we reject $\mathcal{H}_0$. This error is $P_0(A^c)$.
- Error of the second kind, if under the alternative hypothesis $\mathcal{H}_1$ we reject $\mathcal{H}_1$. This error is $P_1(A)$.

Obviously, one decreases the error of the first kind $P_0(A^c)$ if the error of the second kind $P_1(A)$ increases. We can formulate the optimization problem such that minimize the error of the second kind under the condition that the error of the first kind is at most $0 < \alpha < 1$:

$$\min_{A:\, P_0(A^c) \le \alpha} P_1(A). \tag{2.1}$$

In order to solve this problem the Neyman-Pearson Lemma plays an important role.

Theorem 2.1. (NEYMAN, PEARSON (1933)) *Assume that the distributions P_0 and P_1 have densities f_0 and f_1:*

$$P_0(B) = \int_B f_0(\mathbf{x})d\mathbf{x} \quad \textit{and} \quad P_1(B) = \int_B f_1(\mathbf{x})d\mathbf{x}.$$

For a $\gamma > 0$, put

$$A_\gamma = \{\mathbf{x} : f_0(\mathbf{x}) \geq \gamma f_1(\mathbf{x})\}.$$

If for any set A

$$P_0(A^c) \leq P_0(A_\gamma^c)$$

then

$$P_1(A) \geq P_1(A_\gamma).$$

PROOF. Because of the condition of the theorem, we have the following chain of inequalities:

$$\begin{aligned} P_0(A^c) &\leq P_0(A_\gamma^c) \\ P_0(A^c \cap A_\gamma) + P_0(A^c \cap A_\gamma^c) &\leq P_0(A \cap A_\gamma^c) + P_0(A^c \cap A_\gamma^c) \\ \int_{A^c \cap A_\gamma} f_0(x)dx &\leq \int_{A \cap A_\gamma^c} f_0(x)dx. \end{aligned}$$

The definition of A_γ implies that

$$\gamma \int_{A^c \cap A_\gamma} f_1(\mathbf{x})d\mathbf{x} \leq \int_{A^c \cap A_\gamma} f_0(\mathbf{x})d\mathbf{x} \leq \int_{A \cap A_\gamma^c} f_0(\mathbf{x})d\mathbf{x} \leq \gamma \int_{A \cap A_\gamma^c} f_1(\mathbf{x})d\mathbf{x},$$

therefore using the previous chain of derivations in a reverse order we get that

$$P_1(A^c) \leq P_1(A_\gamma^c).$$

□

In Figure 2.1 the blue area illustrates the error of the first kind, while the red area is the error of the second kind.

If for an $0 < \alpha < 1$ there is a $\gamma = \gamma(\alpha)$, which solves the equation

$$P_0(A_\gamma^c) = \alpha,$$

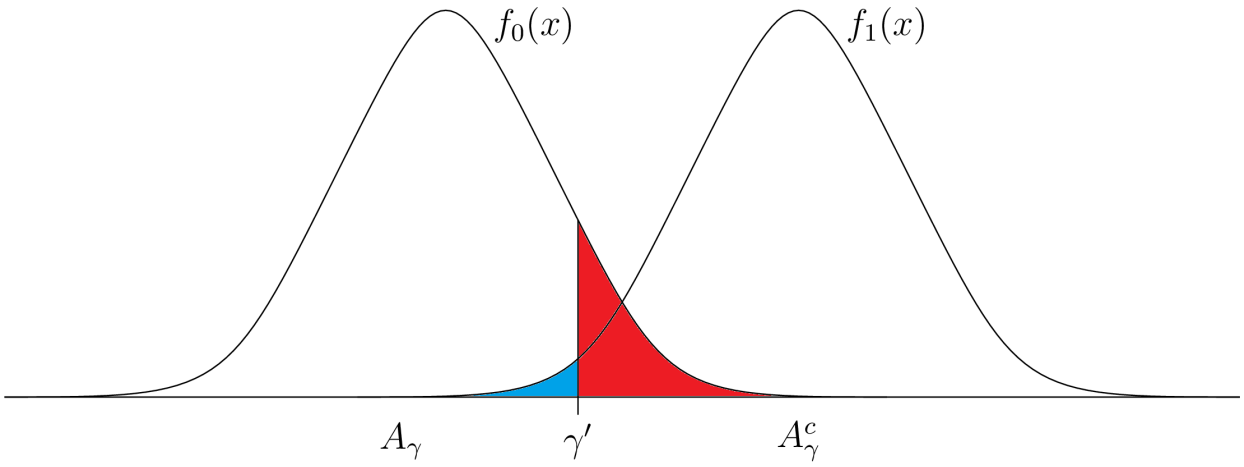

Figure 2.1: Error of the first and second kind.

then the Neyman-Pearson Lemma implies that in order to solve the problem (2.1), it is enough to search for set of form A_γ, i.e.,

$$\min_{A:\, P_0(A^c)\leq\alpha} P_1(A) = \min_{A_\gamma:\, P_0(A_\gamma^c)\leq\alpha} P_1(A_\gamma).$$

Then A_γ is called the *most powerful α-level test.*

Because of the Neyman-Pearson Lemma, we introduce the likelihood ratio statistic

$$T(\mathbf{X}) = \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})},$$

and so the null hypothesis $\mathcal{H}_0$ is accepted if $T(\mathbf{X}) \geq \gamma$.

Example 1. As an illustration of the Neyman-Pearson Lemma, consider the example of an experiment, where the null hypothesis is that the components of $\mathbf{X}$ are i.i.d. normal with mean $m = m_0 > 0$ and with variance σ^2, while under the alternative hypothesis the components of $\mathbf{X}$ are i.i.d. normal with mean $m_1 = 0$ and with the same variance σ^2. Then

$$f_0(\mathbf{x}) = f_0(x_1, \ldots, x_d) = \prod_{i=1}^{d} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - m)^2}{2\sigma^2}} \right)$$

and

$$f_1(\mathbf{x}) = f_1(x_1, \ldots, x_d) = \prod_{i=1}^{d} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}} \right)$$

and

$$\frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} \geq \gamma$$

means that

$$-\sum_{i=1}^{d} \frac{(X_i - m)^2}{2\sigma^2} + \sum_{i=1}^{d} \frac{X_i^2}{2\sigma^2} \geq \ln \gamma,$$

or equivalently,

$$\sum_{i=1}^{d} (2X_i m - m^2) \geq 2\sigma^2 \ln \gamma.$$

This test accepts the null hypothesis if

$$\frac{1}{d} \sum_{i=1}^{d} X_i \geq \frac{2\sigma^2 \ln \gamma / d + m^2}{2m} = \frac{\sigma^2 \ln \gamma}{dm} + \frac{m}{2} =: \gamma'.$$

The test is based on the linear statistic $\sum_{i=1}^{d} X_i/d$, and the question left is how to choose the critical value γ', for which it is an α-level test, i.e., the error of the first kind is α:

$$\mathbb{P}_0 \left\{ \frac{1}{d} \sum_{i=1}^{d} X_i \leq \gamma' \right\} = \alpha.$$

Under the null hypothesis, the distribution of $\frac{1}{d} \sum_{i=1}^{d} X_i$ is normal with mean m and with variance σ^2/d, therefore

$$\mathbb{P}_0 \left\{ \frac{1}{d} \sum_{i=1}^{d} X_i \leq \gamma' \right\} = \Phi \left(\frac{\gamma' - m}{\sigma/\sqrt{d}} \right),$$

where Φ denotes the standard normal distribution function, and so the critical value γ' of an α-level test solves the equation

$$\Phi \left(-\frac{m - \gamma'}{\sigma/\sqrt{d}} \right) = \alpha,$$

i.e.,

$$\gamma' = m - \Phi^{-1}(1 - \alpha)\sigma/\sqrt{d}.$$

REMARK 1. In many situations, when d is large enough, one can refer to the central limit theorem such that the log-likelihood ratio

$$\ln \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}$$

is asymptotically normal. The argument of Example 1 can be extended if under $\mathcal{H}_0$, the log-likelihood ratio is approximately normal with mean m_0 and with variance σ_0^2. Let the test be defined such that it accepts $\mathcal{H}_0$ if

$$\ln \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} \geq \gamma',$$

where

$$\gamma' = m_0 - \Phi^{-1}(1-\alpha)\sigma_0.$$

Then this test is approximately an α-level test.

2.2 ϕ-divergences

In the analysis of repeated observations the divergences between distribution play an important role. Imre Csiszár (1967) introduced the concept of ϕ-divergences. Let $\phi : (0,\infty) \to \mathbb{R}$ be a convex function, extended on $[0,\infty)$ by continuity such that $\phi(1) = 0$. For the probability distributions μ and ν, let λ be a σ-finite dominating measure of μ and ν, for example, $\lambda = \mu + \nu$. Introduce the notations

$$f = \frac{d\mu}{d\lambda}$$

and

$$g = \frac{d\nu}{d\lambda}.$$

Then the *ϕ-divergence* of μ and ν is defined by

$$D_\phi(\mu,\nu) = \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}). \tag{2.2}$$

The Jensen inequality implies the most important property of the ϕ-divergences:

$$D_\phi(\mu,\nu) = \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \geq \phi\left(\int_{\mathbb{R}^d} \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\lambda(d\mathbf{x})\right) = \phi(1) = 0.$$

It means that $D_\phi(\mu,\nu) \geq 0$ and if $\mu = \nu$ then $D_\phi(\mu,\nu) = 0$. If, in addition, ϕ is strictly convex at 1 then $D_\phi(\mu,\nu) = 0$ iff $\mu = \nu$.

Next we show some examples.

- For
$$\phi_1(t) = |t-1|,$$
we get the L_1 *distance*
$$D_{\phi_1}(\mu,\nu) = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| \lambda(d\mathbf{x}).$$

- For
$$\phi_2(t) = (\sqrt{t}-1)^2,$$
we get the *squared Hellinger distance*
$$\begin{aligned} D_{\phi_2}(\mu,\nu) &= \int_{\mathbb{R}^d} \left(\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})}\right)^2 \lambda(d\mathbf{x}) \\ &= 2\left(1 - \int_{\mathbb{R}^d} \sqrt{f(\mathbf{x})g(\mathbf{x})}\lambda(d\mathbf{x})\right). \end{aligned}$$

- For
$$\phi_3(t) = -\ln t,$$
we get the *I-divergence* (called also relative entropy or Kullback-Leibler divergence)
$$I(\mu,\nu) = D_{\phi_3}(\mu,\nu) = \int_{\mathbb{R}^d} \ln\left(\frac{g(\mathbf{x})}{f(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

- For
$$\phi_4(t) = (t-1)^2,$$
we get the χ^2-*divergence*
$$\chi^2(\mu,\nu) = D_{\phi_4}(\mu,\nu) = \int_{\mathbb{R}^d} \frac{(f(\mathbf{x}) - g(\mathbf{x}))^2}{g(\mathbf{x})}\lambda(d\mathbf{x}).$$

An equivalent definition of the ϕ-divergence is
$$D_\phi(\mu,\nu) = \sup_{\mathcal{P}} \sum_j \phi\left(\frac{\mu(A_j)}{\nu(A_j)}\right)\nu(A_j), \tag{2.3}$$
where the supremum is taken over all finite Borel measurable partitions $\mathcal{P} = \{A_j\}$ of $\mathbb{R}^d$.

The main reasoning of this equivalence is that for any partition $\mathcal{P} = \{A_j\}$, the Jensen inequality implies that

$$\begin{aligned} D_\phi(\mu,\nu) &= \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \\ &= \sum_j \int_{A_j} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \\ &= \sum_j \frac{1}{\nu(A_j)} \int_{A_j} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x})\nu(A_j) \\ &\geq \sum_j \phi\left(\frac{1}{\nu(A_j)} \int_{A_j} \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\lambda(d\mathbf{x})\right) \nu(A_j) \\ &= \sum_j \phi\left(\frac{\mu(A_j)}{\nu(A_j)}\right) \nu(A_j). \end{aligned} \tag{2.4}$$

The sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is called nested if any cell $A \in \mathcal{P}_{n+1}$ is a subset of a cell $A' \in \mathcal{P}_n$. Next we show that for nested sequence of partitions

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right) \nu(A) \uparrow .$$

Again, this property is the consequence of the Jensen inequality:

$$\begin{aligned} \sum_{A'\in\mathcal{P}_{n+1}} \phi\left(\frac{\mu(A')}{\nu(A')}\right) \nu(A') &= \sum_{A\in\mathcal{P}_n} \left(\sum_{A'\in\mathcal{P}_{n+1}, A'\subset A} \phi\left(\frac{\mu(A')}{\nu(A')}\right) \nu(A') \right) \\ &= \sum_{A\in\mathcal{P}_n} \left(\sum_{A'\in\mathcal{P}_{n+1}, A'\subset A} \phi\left(\frac{\mu(A')}{\nu(A')}\right) \frac{\nu(A')}{\nu(A)} \right) \nu(A) \\ &\geq \sum_{A\in\mathcal{P}_n} \phi\left(\sum_{A'\in\mathcal{P}_{n+1}, A'\subset A} \frac{\mu(A')}{\nu(A')} \frac{\nu(A')}{\nu(A)} \right) \nu(A) \\ &= \sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right) \nu(A). \end{aligned}$$

It implies that there is a nested sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ such that

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A) \uparrow \sup_{\mathcal{P}_n} \sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A).$$

The sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is called asymptotically fine if for any sphere S centered at the origin

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n, A\cap S\neq 0} diam(A) = 0. \tag{2.5}$$

One can show that if the nested sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is asymptotically fine then

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A) \uparrow \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

This final step will be verified in the particular case of L_1 distance, cf. (4.6). In general, we may introduce a cell wise constant approximation of $\frac{f(\mathbf{x})}{g(\mathbf{x})}$:

$$F_n(\mathbf{x}) := \frac{\mu(A)}{\nu(A)} \text{ if } \mathbf{x}\in A.$$

Thus,

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A) = \int_{\mathbb{R}^d} \phi\left(F_n(\mathbf{x})\right) g(\mathbf{x})\lambda(d\mathbf{x})$$

and

$$F_n(\mathbf{x}) \to \frac{f(\mathbf{x})}{g(\mathbf{x})}$$

for almost all $\mathbf{x} \mod \lambda$ with $g(\mathbf{x}) > 0$ such that

$$\int_{\mathbb{R}^d} \phi\left(F_n(\mathbf{x})\right) g(\mathbf{x})\lambda(d\mathbf{x}) \to \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

2.3 Repeated observations

The error probabilities can be decreased if instead of an observation vector $\mathbf{X}$, we are given n vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ such that under $\mathcal{H}_0$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed (i.i.d.) with distribution P_0, while under $\mathcal{H}_1$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. with distribution P_1. In this case the likelihood ratio statistic is of form

$$T(\mathbf{X}) = \frac{f_0(\mathbf{X}_1)\cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1)\cdot \ldots \cdot f_1(\mathbf{X}_n)}.$$

The Stein Lemma below says that there are tests, for which both the error of the first kind α_n and the error of the second kind β_n tend to 0, if $n \to \infty$.

In order to formulate the Stein Lemma, we remember the *I-divergence*

$$I(P_0, P_1) = D(f_0, f_1) = \int_{\mathbb{R}^d} f_0(\mathbf{x}) \ln \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} d\mathbf{x}. \tag{2.6}$$

.

Theorem 2.2. (CF. CHERNOFF (1952)) *For any $0 < \delta < D(f_0, f_1)$, there is a test such that the error of the first kind*

$$\alpha_n \to 0,$$

and for the error of the second kind

$$\beta_n \leq e^{-n(D(f_0,f_1)-\delta)} \to 0.$$

PROOF. Construct a test such that accept the null hypothesis $\mathcal{H}_0$ if

$$\frac{f_0(\mathbf{X}_1) \cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1) \cdot \ldots \cdot f_1(\mathbf{X}_n)} \geq e^{n(D(f_0,f_1)-\delta)},$$

or equivalently

$$\frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} \geq D(f_0, f_1) - \delta.$$

Under $\mathcal{H}_0$, the strong law of large numbers implies that

$$\frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} \to D(f_0, f_1)$$

almost surely (a.s.), therefore for the error of the first kind α_n, we get that

$$\alpha_n = \mathbb{P}_0 \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} < D(f_0, f_1) - \delta \right\} \to 0.$$

Concerning the error of the second kind β_n we have the following simple bound:

$$\begin{aligned}
& \beta_n \\
= \; & \mathbb{P}_1 \left\{ \frac{f_0(\mathbf{X}_1) \cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1) \cdot \ldots \cdot f_1(\mathbf{X}_n)} \geq e^{n(D(f_0,f_1)-\delta)} \right\} \\
= \; & \int_{\left\{ \frac{f_0(\mathbf{x}_1) \cdot \ldots \cdot f_0(\mathbf{x}_n)}{f_1(\mathbf{x}_1) \cdot \ldots \cdot f_1(\mathbf{x}_n)} \geq e^{n(D(f_0,f_1)-\delta)} \right\}} f_1(\mathbf{x}_1) \cdot \ldots \cdot f_1(\mathbf{x}_n) d\mathbf{x}_1, \ldots, d\mathbf{x}_n \\
\leq \; & e^{-n(D(f_0,f_1)-\delta)} \int_{\left\{ \frac{f_0(\mathbf{x}_1) \cdot \ldots \cdot f_0(\mathbf{x}_n)}{f_1(\mathbf{x}_1) \cdot \ldots \cdot f_1(\mathbf{x}_n)} \geq e^{n(D(f_0,f_1)-\delta)} \right\}} f_0(\mathbf{x}_1) \cdot \ldots \cdot f_0(\mathbf{x}_n) d\mathbf{x}_1, \ldots, d\mathbf{x}_n \\
\leq \; & e^{-n(D(f_0,f_1)-\delta)}.
\end{aligned}$$

□

The critical value of the test in the proof of the Stein Lemma used the I-divergence $D(f_0, f_1)$. Without knowing $D(f_0, f_1)$, the Chernoff Lemma below results in exponential rate of convergence of the errors.

Theorem 2.3. (Chernoff (1952)). *Construct a test such that accept the null hypothesis $\mathcal{H}_0$ if*

$$\frac{f_0(\mathbf{X}_1) \cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1) \cdot \ldots \cdot f_1(\mathbf{X}_n)} \geq 1,$$

or equivalently

$$\sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} \geq 0.$$

(This test is called maximum likelihood test.) Then

$$\alpha_n \leq \left(\inf_{s>0} \int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s} d\mathbf{x} \right)^n$$

and

$$\beta_n \leq \left(\inf_{s>0} \int_{\mathbb{R}^d} f_0(\mathbf{x})^s f_1(\mathbf{x})^{1-s} d\mathbf{x} \right)^n.$$

Proof. Apply the Chernoff bounding technique such that for any $s > 0$ the Markov

inequality implies that

$$\begin{aligned}
\alpha_n &= \mathbb{P}_0\left\{\sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} < 0\right\} \\
&= \mathbb{P}_0\left\{s\sum_{i=1}^{n} \ln \frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)} > 0\right\} \\
&= \mathbb{P}_0\left\{e^{s\sum_{i=1}^{n} \ln \frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}} > 1\right\} \\
&\leq \mathbb{E}_0\left\{e^{s\sum_{i=1}^{n} \ln \frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}}\right\} \\
&= \mathbb{E}_0\left\{\prod_{i=1}^{n}\left(\frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}\right)^s\right\}.
\end{aligned}$$

Under $\mathcal{H}_0$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d., therefore

$$\begin{aligned}
\alpha_n &\leq \mathbb{E}_0\left\{\prod_{i=1}^{n}\left(\frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}\right)^s\right\} \\
&= \prod_{i=1}^{n}\mathbb{E}_0\left\{\left(\frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}\right)^s\right\} \\
&= \mathbb{E}_0\left\{\left(\frac{f_1(\mathbf{X}_1)}{f_0(\mathbf{X}_1)}\right)^s\right\}^n \\
&= \left(\int_{\mathbb{R}^d}\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}\right)^s f_0(\mathbf{x})d\mathbf{x}\right)^n.
\end{aligned}$$

Since $s > 0$ is arbitrary, the first half of the lemma is proved, and the proof of the second half is similar. □

REMARK 2. The Chernoff Lemma results in exponential rate of convergence if

$$\inf_{s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x} < 1$$

and

$$\inf_{s>0}\int_{\mathbb{R}^d} f_0(\mathbf{x})^s f_1(\mathbf{x})^{1-s}d\mathbf{x} < 1.$$

The Cauchy-Schwartz inequality implies that

$$\begin{aligned}\inf_{s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x} &\leq \int_{\mathbb{R}^d} f_1(\mathbf{x})^{1/2} f_0(\mathbf{x})^{1/2}d\mathbf{x} \\ &\leq \sqrt{\int_{\mathbb{R}^d} f_1(\mathbf{x})d\mathbf{x}\int_{\mathbb{R}^d} f_0(\mathbf{x})d\mathbf{x}} \\ &= 1,\end{aligned}$$

with equality in the second inequality if and only if $f_0 = f_1$. Morover, one can check that the function

$$g(s) := \int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x}$$

is convex such that $g(0) = 1$ and $g(1) = 1$, therefore

$$\inf_{s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x} = \inf_{1>s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x}.$$

The quantity

$$He(f_0, f_1) = \int_{\mathbb{R}^d} f_1(\mathbf{x})^{1/2} f_0(\mathbf{x})^{1/2}d\mathbf{x} \tag{2.7}$$

is called *Hellinger integral*. The previous derivations imply that

$$\alpha_n \leq He(f_0, f_1)^n$$

and

$$\beta_n \leq He(f_0, f_1)^n.$$

The squared Hellinger distance $D_{\phi_2}(\mu, \nu)$ was introduced in previous section. One can check that

$$D_{\phi_2}(\mu, \nu) = 2\left(1 - He(f_0, f_1)\right).$$

Chapter 3

Detection

3.1 The detection problem

In this chapter we summarize the basic models, algorithms and results of detection theory. Concerning the details we suggest to visit the books Haykin (1992), (1993), Helstrom (1960), Kang (2008), Levy (2008), Papoulis (1984), Papoulis, Pillai (2002), Skolnik (1980), Trees (1971).

The detection is a hypotheses testing problem with repeated observations such that the null hypothesis $\mathcal{H}_0$ is that in the range of the radar there is no object at a given distance, while the alternative hypothesis $\mathcal{H}_1$ is that there is one.

- The error of the first kind is called false alarm or false detection.
- The error of the second kind is called missed detection.

In the most simple setup the test is based on a single complex valued sample of the reflected signal plus Gaussian noise. Under $\mathcal{H}_0$, the sample is from noise, i.e., it is a complex random variable $N_1 + iN_2$, where N_1 and N_2 are independent, zero mean, Gaussian random variables with variance σ^2. In this case

$$\mathbb{E}\{|N_1 + iN_2|^2\} = \mathbb{E}\{|N_1|^2 + |N_2|^2\} = 2\sigma^2,$$

and

$$\begin{aligned}
\mathbb{V}ar\{|N_1 + iN_2|^2\} &= \mathbb{E}\{(|N_1|^2 + |N_2|^2)^2\} - \mathbb{E}\{|N_1|^2 + |N_2|^2\}^2 \\
&= \mathbb{E}\{|N_1|^4\} + 2\mathbb{E}\{(|N_1|^2|N_2|^2\} + \mathbb{E}\{|N_2|^4\} - (2\sigma^2)^2 \\
&= 3\sigma^4 + 2\sigma^4 + 3\sigma^4 - 4\sigma^4 \\
&= 4\sigma^4,
\end{aligned}$$

where we applied the fact

$$\begin{aligned}
\mathbb{E}\{N_1^4\} &= \sigma^4 \mathbb{E}\left\{\left(\frac{N_1}{\sigma}\right)^4\right\} \\
&= \sigma^4 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^4 e^{-\frac{x^2}{2}} dx \\
&= \sigma^4 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 x e^{-\frac{x^2}{2}} dx \\
&= \sigma^4 \frac{1}{\sqrt{2\pi}} \left(-x^3 e^{-\frac{x^2}{2}} dx \Big|_{-\infty}^{\infty} + 3 \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} dx \right) \\
&= 3\sigma^4
\end{aligned}$$

.

Under $\mathcal{H}_1$, the sample is from signal plus noise, i.e., it is a complex random variable of form

$$N_1 + A\cos\vartheta + i(N_2 + A\sin\vartheta),$$

where A is the amplitude and θ is the phase of the signal. Because of the symmetry of the density of (N_1, N_2), the distribution of $N_1 + A\cos\vartheta + i(N_2 + A\sin\vartheta)$ does not depend on ϑ, therefore we consider only the case $\vartheta = 0$. Then

$$\begin{aligned}
\mathbb{E}\{|N_1 + A + iN_2|^2\} &= \mathbb{E}\{|N_1 + A|^2 + |N_2|^2\} \\
&= 2\sigma^2 + A^2,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{V}ar\{|N_1 + A + iN_2|^2\} \\
&= \mathbb{E}\{(|N_1 + A|^2 + |N_2|^2)^2\} - \mathbb{E}\{|N_1 + A|^2 + |N_2|^2\}^2 \\
&= \mathbb{E}\{|N_1 + A|^4\} + 2\mathbb{E}\{(|N_1 + A|^2|N_2|^2\} + \mathbb{E}\{|N_2|^4\} - (2\sigma^2 + A^2)^2 \\
&= 3\sigma^4 + 6\sigma^2 A^2 + A^4 + 2(\sigma^2 + A^2)\sigma^2 + 3\sigma^4 - 4\sigma^4 - 4\sigma^2 A^2 - A^4 \\
&= 4\sigma^4 + 4\sigma^2 A^2.
\end{aligned}$$

Under $\mathcal{H}_0$, let's calculate the density of the random variable

$$X = \sqrt{N_1^2 + N_2^2}.$$

The density of N_1 and N_2 has the form

$$\frac{1}{\sigma}\varphi\left(\frac{x}{\sigma}\right),$$

where

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is the standard normal density. Then the distribution function of X is as follows:

$$\begin{aligned}
\mathbb{P}\{X \le R\} &= \mathbb{P}\left\{\sqrt{N_1^2 + N_2^2} \le R\right\} \\
&= \int\int_{x^2+y^2 \le R^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} dxdy \\
&= \int\int_{x^2+y^2 \le R^2} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy.
\end{aligned}$$

With polar coordinates, we have that

$$\mathbb{P}\{X \le R\} = \int_{r \le R} \frac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}} 2\pi r dr,$$

which implies the density of X as

$$f_0(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}. \tag{3.1}$$

$f_0(x)$ is called Rayleigh density.

Under $\mathcal{H}_1$, we the distribution function of

$$X = \sqrt{(N_1 + A)^2 + N_2^2}$$

as follows:

$$\begin{aligned}
\mathbb{P}\{X \le R\} &= \mathbb{P}\left\{\sqrt{(N_1 + A)^2 + N_2^2} \le R\right\} \\
&= \int\int_{x^2+y^2 \le R^2} \frac{1}{2\pi\sigma^2} e^{-\frac{(x-A)^2+y^2}{2\sigma^2}} dxdy \\
&= \int\int_{x^2+y^2 \le R^2} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2+A^2-2xA}{2\sigma^2}} dxdy.
\end{aligned}$$

Using the polar coordinates (r, θ), we get that

$$\mathbb{P}\{X \leq R\} = \int_0^R \left(\int_0^{2\pi} \frac{1}{2\pi\sigma^2} e^{-\frac{r^2+A^2-2rA\cos\theta}{2\sigma^2}} d\theta \right) r dr,$$

which implies the Rice density:

$$f_1(x) = \frac{x}{\sigma^2} e^{-\frac{x^2+A^2}{2\sigma^2}} I_0\left(\frac{xA}{\sigma^2}\right), \tag{3.2}$$

where

$$I_0(z) = \frac{1}{\pi}\int_0^{\pi} e^{z\cos\theta} d\theta = \frac{1}{2\pi}\int_0^{2\pi} e^{z\cos\theta} d\theta$$

is the modified Bessel function of zero order. On p. 26 of Skolnik (1980) there is an expansion saying, that for large z

$$I_0(z) = \frac{e^z}{\sqrt{2\pi z}}\left(1 + \frac{1}{8z} + \dots\right) \approx \frac{e^z}{\sqrt{2\pi z}}.$$

We may get this approximation from the second order Taylor expansion of $\cos\theta$:

$$\begin{aligned}
I_0(z) &\approx \frac{1}{\pi}\int_0^{\pi/2} e^{z\cos\theta} d\theta \\
&= e^z \frac{1}{\pi}\int_0^{\pi/2} e^{-z(1-\cos\theta)} d\theta \\
&\approx e^z \frac{1}{\pi}\int_0^{\pi/2} e^{-z\theta^2/2} d\theta \\
&\approx e^z \frac{1}{\pi}\frac{1}{2}\int_{-\infty}^{\infty} e^{-\frac{\theta^2}{2/z}} d\theta \\
&= e^z \frac{1}{\pi}\frac{1}{2}\sqrt{2\pi/z} \\
&= \frac{e^z}{\sqrt{2\pi z}}.
\end{aligned}$$

(3.1) and (3.2) imply that the likelihood ratio has the form

$$\frac{f_0(x)}{f_1(x)} = \frac{e^{\frac{A^2}{2\sigma^2}}}{I_0\left(\frac{xA}{\sigma^2}\right)} \approx e^{\frac{A^2}{2\sigma^2}-\frac{xA}{\sigma^2}} \sqrt{2\pi\frac{xA}{\sigma^2}}. \tag{3.3}$$

3.2 Two non-coherent detection algorithms

Assume that the radar repeats sending the signal n times, and after compressed filtering the received signal is sampled such that the number of samples is denoted by N. The samples of the k-th received signal are collected in a cluster vector

$$\mathbf{u}_k^T = \begin{pmatrix} y_{k,1} & \dots & y_{k,j_0} & \dots & y_{k,N} \end{pmatrix},$$

$(k = 0, 1, \dots, n-1)$, while the set of cluster vectors forms the matrix

$$\mathbf{Y} = \begin{pmatrix} y_{0,1} & \dots & y_{0,j_0} & \dots & y_{0,N} \\ y_{1,1} & \dots & y_{1,j_0} & \dots & y_{1,N} \\ \vdots & \dots & \vdots & \ddots & \vdots \\ y_{n-1,1} & \dots & y_{n-1,j_0} & \dots & y_{n-1,N} \end{pmatrix}. \tag{3.4}$$

Algorithm 1. Let $\mathcal{H}_0$ be the hypothesis, that there no object at a distance corresponding to the index j_0. It means that the task of detection is combined with the distance estimation. The test is based on

$$|y_{0,j_0}|^2, |y_{1,j_0}|^2, \dots, |y_{n-1,j_0}|^2$$

such that from the matrix $\mathbf{Y}$ calculate

$$z_j^2 = \sum_{k=0}^{n-1} |y_{k,j}|^2,$$

$(j = 1, \dots, N)$. Under $\mathcal{H}_0$, the central limit theorem (CLT) implies that $z_{j_0}^2$ is approximately Gaussian distributed with mean $n2\sigma^2$ and with variance $n4\sigma^4$, while under $\mathcal{H}_1$ the mean is $n(2\sigma^2 + A^2)$ and the variance is $n(4\sigma^4 + 4\sigma^2 A^2)$.

We may choose a threshold

$$\gamma' = 2\sigma^2 + A^2/2.$$

The amplitude A is unknown, it depens on many factors like the size, velocity and profile of the object, meteorology, etc. Therefore γ' cannot be chosen in this way. Instead a value A_{min} is introduced as follows. Let

$$SNR := \frac{A^2}{\sigma^2}$$

be the signal-to-noise ratio. Assume that we are given a minimum acceptable value of signal-to-noise ratio SNR_{min}, for example,

$$SNR_{min} = 1.$$

Let's estimate the variance σ^2 by

$$2\hat{\sigma}^2 = \frac{1}{nN}\sum_{k=0}^{n-1}\sum_{j=1}^{N}|y_{k,j}|^2$$

Under $\mathcal{H}_0$, $\hat{\sigma}^2$ is an unbiased estimate of σ^2, while under $\mathcal{H}_1$, $\hat{\sigma}^2$ overestimates σ^2 such that $2\hat{\sigma}^2 \approx 2\sigma^2 + A^2/N$. Put

$$A_{min} = \sqrt{SNR_{min}}\hat{\sigma}$$

and

$$\gamma' = 2\hat{\sigma}^2 + A_{min}^2/2.$$

Accept $\mathcal{H}_0$, if

$$\frac{1}{n}\sum_{k=0}^{n-1}|y_{k,j_0}|^2 \leq \gamma',$$

reject otherwise.

Again, the CLT implies that the probability of false alarm is approximately

$$\begin{aligned}
\alpha_n &\approx \Phi\left(-\frac{n\gamma' - n2\sigma^2}{\sqrt{n4\sigma^4}}\right) \\
&= \Phi\left(-\sqrt{n}\frac{\gamma' - 2\sigma^2}{2\sqrt{\sigma^4}}\right) \\
&\approx \Phi\left(-\sqrt{n}\frac{(A_{min}^2/2)/\sigma^2}{2}\right) \\
&= \Phi\left(-\sqrt{n}\frac{SNR_{min}}{4}\right),
\end{aligned}$$

while the probability of missed detection is approximately

$$\begin{aligned}\beta_n &\approx \Phi\left(\frac{n\gamma' - n(2\sigma^2 + A^2)}{\sqrt{n(4\sigma^4 + 4\sigma^2 A^2)}}\right) \\ &= \Phi\left(-\sqrt{n}\frac{2\sigma^2 + A^2 - \gamma'}{2\sqrt{\sigma^4 + \sigma^2 A^2}}\right) \\ &\approx \Phi\left(-\sqrt{n}\frac{(A^2 - A_{min}^2/2)/\sigma^2}{2\sqrt{1 + A^2/\sigma^2}}\right) \\ &= \Phi\left(-\sqrt{n}\frac{SNR - SNR_{min}/2}{2\sqrt{1 + SNR}}\right).\end{aligned}$$

Evaluating both probabilities, it is useful to have the bounds

$$\frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t}\left(1 - \frac{1}{t^2}\right) \leq \Phi(-t) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t},$$

($t > 0$, cf. p. 179 in Feller (1968)). Because of $\Phi(-t) \leq 1/2$, the upper bound implies that

$$\Phi(-t) \leq e^{-t^2/2}.$$

Thus,

$$\alpha_n \leq e^{-n\frac{SNR_{min}^2}{8}},$$

and

$$\beta_n \leq e^{-n\frac{(SNR - SNR_{min}/2)^2}{4(1+SNR)}}.$$

Therefore both error probabilities tends to zero exponentially fast. We can illustrate the upper bound by the case of $n = 256$ and $SNR_{min} = 1$:

$$\alpha_n \leq 10^{-14}.$$

Algorithm 2. Let $\mathcal{H}_0$ be as before. The test is based on

$$|y_{0,j_0}|, |y_{1,j_0}|, \ldots, |y_{n-1,j_0}|.$$

Under $\mathcal{H}_0$, the density of $|y_{k,j_0}|$ is according to (3.1), while under $\mathcal{H}_1$, the density of $|y_{k,j_0}|$ is given by (3.2). Therefore the maximum likelihood test accept $\mathcal{H}_0$,

$$\sum_{k=0}^{n-1} \ln \frac{f_0(|y_{k,j_0}|)}{f_1(|y_{k,j_0}|)} \geq 0.$$

Because of (3.3), the approximately maximum likelihood test accept $\mathcal{H}_0$, if

$$\sum_{k=0}^{n-1}\left(\frac{A^2}{2\sigma^2}-\frac{|y_{k,j_0}|A}{\sigma^2}\right)\geq 0,$$

or equivalently

$$\sum_{k=0}^{n-1}\left(\frac{A}{2}-|y_{k,j_0}|\right)\geq 0.$$

If A is replaced by A_{min}, then we get the test, which accepts $\mathcal{H}_0$-t, if

$$\sum_{k=0}^{n-1}\left(\frac{A_{min}}{2}-|y_{k,j_0}|\right)\geq 0,$$

or equivalently

$$\frac{1}{n}\sum_{k=0}^{n-1}|y_{k,j_0}|\leq\frac{A_{min}}{2}.$$

The modified, approximately maximum likelihood test accepts $\mathcal{H}_0$, if

$$\frac{1}{n}\sum_{k=0}^{n-1}|y_{k,j_0}|\leq\frac{A_{min}}{2}+\hat{\sigma}\sqrt{\frac{\pi}{2}},$$

and reject otherwise.

Let's calculate the error probabilities. Put

$$u_{j_0}=\sum_{k=0}^{n-1}|y_{k,j_0}|.$$

From the formula of Rayleigh density one can derive, that under $\mathcal{H}_0$

$$\mathbb{E}\{|y_{k,j_0}|\}=\sigma\sqrt{\frac{\pi}{2}}$$

and

$$\mathbb{V}ar\{|y_{k,j_0}|\}=\sigma^2\left(2-\frac{\pi}{2}\right),$$

because using the notation of (3.1) we get

$$\begin{aligned}\mathbb{E}\{|y_{k,j_0}|\} &= \int_0^\infty x f_0(x) dx \\ &= \int_0^\infty \frac{x^2}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{1}{2}\frac{\sqrt{2\pi}}{\sigma}\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^\infty x^2 e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \sigma\sqrt{\frac{\pi}{2}},\end{aligned}$$

and

$$\mathbb{V}ar\{|y_{k,j_0}|\} = \mathbb{E}\{|y_{k,j_0}|^2\} - \mathbb{E}\{|y_{k,j_0}|\}^2 = \sigma^2\left(2 - \frac{\pi}{2}\right).$$

The CLT implies that the distribution of u_{j_0} is approximately normal with mean $n\sigma\sqrt{\pi/2}$ and with variance $n\sigma^2(2-\pi/2)$.

Then the false alarm probability is approximately equal to

$$\begin{aligned}\alpha_n &\approx \Phi\left(-\frac{nA_{min}/2 + n\sigma\sqrt{\pi/2} - n\sigma\sqrt{\pi/2}}{\sqrt{n\sigma^2(2-\pi/2)}}\right) \\ &= \Phi\left(-\sqrt{n}\frac{A_{min}/2}{\sqrt{2-\pi/2}\sigma}\right) \\ &= \Phi\left(-\sqrt{n}\frac{\sqrt{SNR_{min}}/2}{\sqrt{2-\pi/2}}\right) \\ &\leq e^{-n\frac{(\sqrt{SNR_{min}}/2)^2}{2(2-\pi/2)}} \\ &\leq e^{-nSNR_{min}/4}.\end{aligned}$$

Under $\mathcal{H}_1$, the Jensen inequality implies

$$A \leq \mathbb{E}\{|y_{k,j_0}|\},$$

while from

$$|y_{k,j_0}| = \sqrt{(N_1 + A)^2 + N_2^2} \leq A + \sqrt{N_1^2 + N_2^2}$$

one gets

$$\mathbb{E}\{|y_{k,j_0}|\} \leq A + \sigma\sqrt{\frac{\pi}{2}}.$$

Moreover,

$$\mathbb{V}ar(|y_{k,j_0}|) = \mathbb{E}\{|y_{k,j_0}|^2\} - \mathbb{E}\{|y_{k,j_0}|\}^2 \leq 2\sigma^2 + A^2 - A^2 = 2\sigma^2.$$

Therefore the CLT implies that the distribution function of u_{j_0} can be lower bounded by the normal distribution function with mean nA and with variance $n2\sigma^2$.

In this way we have an approximate upper bound on the missed detection probability:

$$\begin{aligned}
\beta_n &\leq \Phi\left(\frac{nA_{min}/2 + n\sigma\sqrt{\pi/2} - nA}{\sqrt{n2\sigma^2}}\right) \\
&= \Phi\left(-\sqrt{n}\frac{A - A_{min}/2 - \sigma\sqrt{\pi/2}}{\sqrt{2}\sigma}\right) \\
&= \Phi\left(-\sqrt{n}\frac{\sqrt{SNR} - \sqrt{SNR_{min}}/2 - \sqrt{\pi/2}}{\sqrt{2}}\right) \\
&\leq e^{-n\frac{(\sqrt{SNR} - \sqrt{SNR_{min}}/2 - \sqrt{\pi/2})^2}{4}},
\end{aligned}$$

provided, that $\sqrt{SNR} - \sqrt{SNR_{min}}/2 - \sqrt{\pi/2} \geq 0$.

3.3 DFT based detection

Introduce the matrix

$$\mathbf{RV} = \begin{pmatrix} Y_{0,1} & \dots & Y_{0,j} & \dots & Y_{0,N} \\ Y_{1,1} & \dots & Y_{1,j} & \dots & Y_{1,N} \\ \vdots & \dots & \vdots & \ddots & \vdots \\ Y_{n-1,1} & \dots & Y_{n-1,j} & \dots & Y_{n-1,N} \end{pmatrix} \tag{3.5}$$

such that

$$Y_{m,j} = \sum_{k=0}^{n-1} y_{k,j} e^{-i2\pi\frac{m}{n}\cdot k}.$$

Notice that the j-th column of $\mathbf{RV}$ is the DFT of the j-th column of $\mathbf{Y}$, and the pair (m, j) corresponds to a velocity, distance cell.

Algorithm 3. In contrast to the previous section, here the hypothesis $\mathcal{H}_0$ is, that there is no object at a distance corresponding to the index j_0, which moves with velocity

corresponding to the index m_0. It means that the task of detection is combined with the distance and velocity estimation. The test is based on

$$|Y_{m_0,j_0}|.$$

Under $\mathcal{H}_0$, Y_{m_0,j_0} is a sum of n independent, complex valued, Gaussian, zero mean random variables with variance σ^2. Therefore $|Y_{m_0,j_0}|$ has a Rayleigh density:

$$f_{0,n}(x) = \frac{x}{n\sigma^2} e^{-\frac{x^2}{2n\sigma^2}}.$$

Earlier A denoted the amplitude, which depends of on the velocity because of Doppler effect. Under $\mathcal{H}_1$, because of matched DFT the amplitude is almost independent of the velocity. In the sequel, this amplitude is denoted by $\bar{A}$. Under $\mathcal{H}_1$, $|Y_{m_0,j_0}|$ has Rice density:

$$f_{1,n}(x) = \frac{x}{n\sigma^2} e^{-\frac{x^2+n^2\bar{A}^2}{2n\sigma^2}} I_0\left(\frac{xn\bar{A}}{n\sigma^2}\right).$$

Thus, the likelihood ratio is

$$\frac{f_{0,n}(x)}{f_{1,n}(x)} = \frac{e^{\frac{n\bar{A}^2}{2\sigma^2}}}{I_0\left(\frac{x\bar{A}}{\sigma^2}\right)} \approx e^{\frac{n\bar{A}^2}{2\sigma^2} - \frac{x\bar{A}}{\sigma^2}} \sqrt{2\pi \frac{x\bar{A}}{\sigma^2}} \approx e^{\frac{n\bar{A}^2}{2\sigma^2} - \frac{x\bar{A}}{\sigma^2}}.$$

According to the approximately maximum likelihood test, we accept $\mathcal{H}_0$, if

$$nA_{min}/2 \geq |Y_{m_0,j_0}|,$$

and reject otherwise.

Then the false alarm probability is

$$\begin{aligned}
\alpha_n &= \int_{nA_{min}/2}^{\infty} f_{0,n}(x)dx \\
&= \int_{nA_{min}/2}^{\infty} \frac{x}{n\sigma^2} e^{-\frac{x^2}{2n\sigma^2}} dx \\
&= e^{-\frac{(nA_{min}/2)^2}{2n\sigma^2}} \\
&= e^{-n\frac{SNR_{min}}{8}},
\end{aligned}$$

while we can upper bound the missed detection probability

$$\beta_n = \int_0^{nA_{min}/2} f_{1,n}(x)dx.$$

The density $f_{1,n}(x)$ is the density of the random variable $\sqrt{(\sqrt{n}N_1 + n\bar{A})^2 + (\sqrt{n}N_2)^2}$. Thus,

$$\begin{aligned}
\beta_n &= \mathbb{P}\{\sqrt{(\sqrt{n}N_1 + n\bar{A})^2 + (\sqrt{n}N_2)^2} \leq nA_{min}/2\} \\
&\leq \mathbb{P}\{|\sqrt{n}N_1 + n\bar{A}| \leq nA_{min}/2\} \\
&\leq \mathbb{P}\{\sqrt{n}N_1 + n\bar{A} \leq nA_{min}/2\} \\
&= \Phi\left(\frac{n(A_{min}/2 - \bar{A})}{\sqrt{n}\sigma}\right) \\
&= \Phi\left(-\sqrt{n}\frac{\bar{A} - A_{min}/2}{\sigma}\right) \\
&\leq e^{-n\frac{(\bar{A}-A_{min}/2)^2}{2\sigma^2}} \\
&= e^{-n\frac{(\sqrt{\mathbb{S}NR}-\sqrt{SNR_{min}}/2)^2}{2}},
\end{aligned}$$

where

$$\mathbb{S}NR = \frac{\bar{A}^2}{\sigma^2}$$

is a larger signal-to-noise ratio than in the previous section.

3.4 Robust detection

In real life problems $\mathcal{H}_0$ and $\mathcal{H}_1$ are composit hypotheses, i.e., both consit of many densities. For example, $\mathcal{H}_0$ consists of many Rayleigh densities, and $\mathcal{H}_1$ consits of many Rice densities. Moreover, in practice the classical detection model is not appropriate, which means that for the received signal, the noise is not Gaussian due to the background or to fading. In the model of robust detection, let $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ be nominal densities, by which we define the two hypotheses such that the true density of the observation belongs to a neighborhood of the nominal densities.

If $f(\mathbf{x})$ and $g(\mathbf{x})$ are densities, then introduce their L_1 distance:

$$\|f - g\| = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})|\, d\mathbf{x}.$$

The L_1 distance is an important quantity, because it results in an upper bound on the difference of probabilities, i.e., the Scheffé Theorem below shows that the total variation is the half of the L_1 distance of the corresponding densities.

Theorem 3.1. *(*SCHEFFÉ (1947)*) If μ and ν are absolutely continuous probability distributions with densities f and g, respectively, then*

$$\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} = 2\, V(\mu, \nu).$$

(The quantity

$$L_1(f, g) = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} \tag{3.6}$$

*is called L_1-*distance*.)*

PROOF. Note that

$$\begin{aligned} V(\mu, \nu) &= \sup_A |\mu(A) - \nu(A)| \\ &= \sup_A \left| \int_A f(\mathbf{x}) d\mathbf{x} - \int_A g(\mathbf{x}) d\mathbf{x} \right| \\ &= \sup_A \left| \int_A (f(\mathbf{x}) - g(\mathbf{x})) d\mathbf{x} \right| \\ &= \int_{f(\mathbf{x}) > g(\mathbf{x})} (f(\mathbf{x}) - g(\mathbf{x})) d\mathbf{x} \\ &= \int_{g(\mathbf{x}) > f(\mathbf{x})} (g(\mathbf{x}) - f(\mathbf{x})) d\mathbf{x} \\ &= \frac{1}{2} \int |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x}. \end{aligned}$$

□

The Scheffé theorem means that, for any set B, we have

$$\left| \int_B f(\mathbf{x}) d\mathbf{x} - \int_B g(\mathbf{x}) d\mathbf{x} \right| \leq \int_{\{\mathbf{x}; f(\mathbf{x}) - g(\mathbf{x}) > 0\}} (f(\mathbf{x}) - g(\mathbf{x})) d\mathbf{x} = \frac{1}{2} \|f - g\|.$$

Let the repeated observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent random vectors with the common density $f(\mathbf{x})$. Under $\mathcal{H}_j$, the density $f(\mathbf{x})$ is a distortion of the nominal density $f_j(\mathbf{x})$, $j = 0, 1$. Formally,

$$\mathcal{H}_0 = \{f(\mathbf{x}) : \|f - f_0\| < \Delta\}, \tag{3.7}$$

and

$$\mathcal{H}_1 = \{f(\mathbf{x}) : \|f - f_1\| < \Delta\}, \tag{3.8}$$

where

$$\Delta := (1/2)\|f_0 - f_1\|.$$

Let B^* be the acceptance set of the maximum likelihood test for the nominal densities:

$$B^* = \{\mathbf{x} : f_0(\mathbf{x}) > f_1(\mathbf{x})\}.$$

Accept $\mathcal{H}_0$, if

$$\mu_n(B^*) \geq \frac{\int_{B^*} f_0(\mathbf{x})d\mathbf{x} + \int_{B^*} f_1(\mathbf{x})d\mathbf{x}}{2}, \tag{3.9}$$

and reject otherwise, where

$$\mu_n(B^*) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\{\mathbf{X}_j \in B^*\}} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\{f_0(\mathbf{X}_j) > f_1(\mathbf{X}_j)\}}.$$

Notice that this test is based on a non-linear statistic, i.e., it is a majority voting test.

Theorem 3.2. (Devroye, Györfi, Lugosi (2002), Györfi, Walk (2014), Biglieri, Györfi (2014).)

$$\alpha_n \leq e^{-n(\Delta - \|f - f_0\|)^2/2}$$

and

$$\beta_n \leq e^{-n(\Delta - \|f - f_1\|)^2/2}.$$

Proof. Put

$$\epsilon = \Delta - \|f - f_0\| > 0.$$

Under $\mathcal{H}_0$, the Scheffé theorem implies

$$\begin{aligned} 2\left(\int_{B^*} f_0(\mathbf{x})d\mathbf{x} - \int_{B^*} f(\mathbf{x})d\mathbf{x}\right) &\leq \|f - f_0\| \\ &= \Delta - \epsilon \\ &= \frac{1}{2}\|f_0 - f_1\| - \epsilon \\ &= \int_{B^*} f_0(\mathbf{x})d\mathbf{x} - \int_{B^*} f_1(\mathbf{x})d\mathbf{x} - \epsilon, \end{aligned}$$

From which we get

$$\int_{B^*} f(\mathbf{x})d\mathbf{x} \geq \frac{\int_{B^*} f_0(\mathbf{x})d\mathbf{x} + \int_{B^*} f_1(\mathbf{x})d\mathbf{x}}{2} + \epsilon/2. \tag{3.10}$$

The Hoeffding inequality (Lemma 1.4) says, that for binary valued and i.i.d. random variables $Z_1, \ldots, Z_n$, one has for $t > 0$,

$$\mathbb{P}\left\{\mathbb{P}\{Z_1 = 1\} - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq t\right\} \leq \exp\left(-2nt^2\right).$$

(3.10) and the Hoeffding inequality imply, that

$$\begin{aligned}
\alpha_n &= \mathbb{P}\left\{\mu_n(B^*) < \frac{\int_{B^*} f_0(\mathbf{x})d\mathbf{x} + \int_{B^*} f_1(\mathbf{x})d\mathbf{x}}{2}\right\} \\
&\leq \mathbb{P}\left\{\int_{B^*} f(\mathbf{x})d\mathbf{x} - \mu_n(B^*) > \epsilon/2\right\} \\
&\leq \mathbb{P}\left\{\mathbb{P}\{f_0(\mathbf{X}_1) > f_1(\mathbf{X}_1)\} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\{f_0(\mathbf{X}_j) > f_1(\mathbf{X}_j)\}} > \epsilon/2\right\} \\
&\leq e^{-n\epsilon^2/2}.
\end{aligned}$$

The proof of the second half of the theorem is similar. □

Notice, that for the threshold in (3.9), in practice we usually one has

$$\int_{B^*} f_0(\mathbf{x})d\mathbf{x} + \int_{B^*} f_1(\mathbf{x})d\mathbf{x} = \int_{B^*} f_0(\mathbf{x})d\mathbf{x} + 1 - \int_{B^{*c}} f_1(\mathbf{x})d\mathbf{x} \approx 1$$

Thus, we a modification of (3.9) as

$$\mu_n(B^*) \geq \frac{1}{2}. \tag{3.11}$$

Algorithm 4. Similarly to Algorithm 2, the test is based on

$$|y_{0,j_0}|, |y_{1,j_0}|, \ldots, |y_{n-1,j_0}|.$$

However, here we have composite hypotheses, and a nonlinear test statistic is applied. For $0 < \epsilon < 1/2$ and for the threshold A_{min}, put

$$\mathcal{H}_0 = \left\{f(x) : \int_0^{A_{min}} f(x)dx - 1/2 \geq \epsilon\right\}$$

and

$$\mathcal{H}_1 = \left\{ f(x) : 1/2 - \int_0^{A_{min}} f(x)dx \geq \epsilon \right\}.$$

The approximately robust detection accepts $\mathcal{H}_0$, if

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} \geq \frac{1}{2},$$

which means, that the ratio of sample's absolute values is less than $1/2$.

The Hoeffding inequality implies upper bounds for the error probabilities:

$$\begin{aligned}
\alpha_n &= \mathbb{P}_0 \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} < 1/2 \right\} \\
&= \mathbb{P}_0 \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} - \mathbb{P}_0\{A_{min} > |y_{0,j_0}|\} < 1/2 - \mathbb{P}_0\{A_{min} > |y_{0,j_0}|\} \right\} \\
&\leq \mathbb{P}_0 \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} - \mathbb{P}_0\{A_{min} > |y_{0,j_0}|\} < -\epsilon \right\} \\
&\leq e^{-2n\epsilon^2}
\end{aligned}$$

and

$$\begin{aligned}
\beta_n &= \mathbb{P}_1 \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} \geq 1/2 \right\} \\
&= \mathbb{P}_1 \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} - \mathbb{P}_1\{A_{min} > |y_{0,j_0}|\} \geq 1/2 - \mathbb{P}_1\{A_{min} > |y_{0,j_0}|\} \right\} \\
&\leq \mathbb{P}_1 \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{I}_{\{A_{min} > |y_{k,j_0}|\}} - \mathbb{P}_1\{A_{min} > |y_{0,j_0}|\} \geq \epsilon \right\} \\
&\leq e^{-2n\epsilon^2}.
\end{aligned}$$

Evaluate these results in the special case, when $\mathcal{H}_0$ is a set of Rayleigh densities, while $\mathcal{H}_1$ consists of Rice densities. Define the hypotheses by

$$\mathcal{H}_0 = \left\{ \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} : \int_0^{A_{min}} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx - 1/2 \geq \epsilon \right\}$$

and

$$\mathcal{H}_1 = \left\{ \frac{x}{\sigma^2} e^{-\frac{x^2+A^2}{2\sigma^2}} I_0\left(\frac{xA}{\sigma^2}\right) : 1/2 - \int_0^{A_{min}} \frac{x}{\sigma^2} e^{-\frac{x^2+A^2}{2\sigma^2}} I_0\left(\frac{xA}{\sigma^2}\right) dx \geq \epsilon \right\}.$$

We have that

$$\mathbb{P}_0\{A_{min} > |y_{0,j_0}|\} = 1 - \int_{A_{min}}^{\infty} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx = 1 - e^{-\frac{A_{min}^2}{2\sigma^2}}.$$

For the hypothesis $\mathcal{H}_0$, we need

$$1/2 - e^{-SNR_{min}/2} \geq \epsilon,$$

or equivalently $SNR_{min} \geq -2\ln(1/2-\epsilon) \approx 2\ln 2 \approx 1.4$. Denote by

$$x^+ = \max\{x, 0\}$$

positive part of x. Then

$$\alpha_n \leq e^{-n\left(\left[1-2e^{-SNR_{min}/2}\right]^+\right)^2/2}.$$

For the hypothesis $\mathcal{H}_0$, we have that

$$\begin{aligned}
\mathbb{P}_1\{A_{min} \geq |y_{0,j_0}|\} &= \mathbb{P}\{\sqrt{(N_1+A)^2 + (N_2)^2} \leq A_{min}\} \\
&\leq \mathbb{P}\{|N_1 + A| \leq A_{min}\} \\
&\leq \mathbb{P}\{N_1 + A \leq A_{min}\} \\
&= \Phi\left(\frac{A_{min} - A}{\sigma}\right) \\
&= \Phi\left(-\frac{A - A_{min}}{\sigma}\right) \\
&\leq e^{-\frac{(A-A_{min})^2}{2\sigma^2}} \\
&= e^{-\frac{(\sqrt{SNR}-\sqrt{SNR_{min}})^2}{2}}.
\end{aligned}$$

Therefore we need

$$1/2 - \int_0^{A_{min}} \frac{x}{\sigma^2} e^{-\frac{x^2+A^2}{2\sigma^2}} I_0\left(\frac{xA}{\sigma^2}\right) dx \geq \epsilon,$$

which is satisfied, if

$$1/2 - e^{-\frac{(\sqrt{SNR}-\sqrt{SNR_{min}})^2}{2}} \geq \epsilon,$$

or equivalently

$$\sqrt{SNR} \geq \sqrt{SNR_{min}} + \sqrt{-2\ln(1/2-\epsilon)} \approx \sqrt{SNR_{min}} + \sqrt{2\ln 2} \approx \sqrt{SNR_{min}} + 1.2$$

Thus,

$$\beta_n \leq e^{-n\left(\left[1-2e^{-\frac{(\sqrt{SNR}-\sqrt{SNR_{min}})^2}{2}}\right]^+\right)^2/2}.$$

3.5 Comparison of the algorithms

For $n = 256$ and for $e^{256} \approx 10^{111}$, Table 3.1 shows the formulas of the algorithms. In addition, we may choose $SNR_{min} = 2$ and $SNR = \mathbb{S}NR = 8$. Then Table 3.2 contains the error probabilities. The coherent signal-to-noise ratio $\mathbb{S}NR$ is much larger than the non-coherent SNR, therefore Algorithm 3 is the best, while Algorithms 1 and 2 have approximately the same good performance. Algorithm 4 is much weaker. However, it works even in the case, when the additive Gaussian noise condition is not satisfied. One can decrease its error probabilities from 10^{-4} to 10^{-8} by doubling n.

	false alarm	missed detection
Algorithm 1	$10^{-14\cdot SNR_{min}^2}$	$10^{-28\frac{(SNR-SNR_{min}/2)^2}{1+SNR}}$
Algorithm 2	$10^{-28\cdot SNR_{min}}$	$10^{-28(\sqrt{SNR}-\sqrt{SNR_{min}}/2-\sqrt{\pi/2})^2}$
Algorithm 3	$10^{-14\cdot SNR_{min}}$	$10^{-56(\sqrt{\mathbb{S}NR}-\sqrt{SNR_{min}}/2)^2}$
Algorithm 4	$10^{-56\left(\left[1-2e^{-SNR_{min}/2}\right]^+\right)^2}$	$10^{-56\left(\left[1-2e^{-\frac{(\sqrt{SNR}-\sqrt{SNR_{min}})^2}{2}}\right]^+\right)^2}$

Table 3.1: The formulas of the error probabilities.

	false alarm	missed detection
Algorithm 1	10^{-56}	10^{-152}
Algorithm 2	10^{-56}	10^{-21}
Algorithm 3	10^{-28}	10^{-251}
Algorithm 4	10^{-4}	10^{-4}

Table 3.2: The error probabilities of the algorithms.

Chapter 4

Testing Simple versus Composite Hypotheses

4.1 Total variation and I-divergence

If μ and ν are probability distributions on $\mathbb{R}^d$ ($d \geq 1$), then the *total variation distance* between μ and ν was defined by

$$V(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|,$$

where the supremum is taken over all Borel sets A. According to the Scheffé Theorem (Theorem 3.1), the total variation is the half of the L_1 distance of the corresponding densities.

The following inequality, called Pinsker's inequality, gives an upper bound to the total variation in terms of I-divergence:

Theorem 4.1. (Csiszár (1967), Kullback (1967) and Kemperman (1969))

$$2\{V(\mu, \nu)\}^2 \leq I(\mu, \nu). \tag{4.1}$$

Proof. Applying the notations of the proof of the Scheffé Theorem (Theorem 3.1), put

$$A^* = \{f > g\},$$

then the Scheffé Theorem implies that

$$V(\mu, \nu) = \mu(A^*) - \nu(A^*).$$

Moreover, from (2.4) we get that

$$I(\mu,\nu) \geq \mu(A^*) \ln \frac{\mu(A^*)}{\nu(A^*)} + (1-\mu(A^*)) \ln \frac{1-\mu(A^*)}{1-\nu(A^*)}$$

Introduce the notations

$$q = \nu(A^*) \text{ and } p = \mu(A^*) > q,$$

and

$$h_p(q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

then we have to prove that

$$2(p-q)^2 \leq h_p(q),$$

which follows from the facts on the derivative:

$$\begin{aligned} \frac{d}{dq}(h_p(q) - 2(p-q)^2) &= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) \\ &= -\frac{p-q}{q(1-q)} + 4(p-q) \\ &\leq 0. \end{aligned}$$

□

4.2 Large deviation of L_1 distance

Consider the sample of $\mathbb{R}^d$-valued random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with *i.i.d.* components such that the common distribution is denoted by ν. For a fixed distribution μ, we consider the problem of testing hypotheses

$$\mathcal{H}_0 : \nu = \mu \text{ versus } \mathcal{H}_1 : \nu \neq \mu$$

by means of test statistics $T_n = T_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

For testing a simple hypothesis $\mathcal{H}_0$ that the distribution of the sample is μ, versus a composite alternative, Györfi and van der Meulen (1990) introduced a related goodness of fit test statistic L_n defined as

$$L_n = \sum_{j=1}^{m_n} |\mu_n(A_{n,j}) - \mu(A_{n,j})|,$$

where μ_n denotes the empirical measure associated with the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, so that

$$\mu_n(A) = \frac{\#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}}{n}$$

for any Borel subset A, and $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ is a finite partition of $\mathbb{R}^d$.

Next we characterize the large deviation properties of L_n:

Theorem 4.2. (Beirlant, Devroye, Györfi and Vajda (2001)). *Assume that*

$$\lim_{n\to\infty} \max_j \mu(A_{n,j}) = 0 \tag{4.2}$$

and

$$\lim_{n\to\infty} \frac{m_n \ln n}{n} = 0. \tag{4.3}$$

Then for all $0 < \epsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\{L_n > \epsilon\} = -g_L(\epsilon), \tag{4.4}$$

where

$$g_L(\epsilon) = \inf_{0<p<1-\epsilon/2} \left(p \ln \frac{p}{p+\epsilon/2} + (1-p) \ln \frac{1-p}{1-p-\epsilon/2} \right). \tag{4.5}$$

Biau and Györfi (2005) provided an alternative derivation of $g_L(\varepsilon)$ and non-asymptotic upper bound.

Theorem 4.3. (Biau and Györfi (2005)). *For any* $\epsilon > 0$,

$$\mathbb{P}\{L_n > \epsilon\} \le 2^{m_n} e^{-n\epsilon^2/2}.$$

Proof. By Scheffé's theorem for partitions

$$L_n = \sum_{A\in\mathcal{P}_n} |\mu_n(A) - \mu(A)| = 2 \max_{A\in\sigma(\mathcal{P}_n)} (\mu_n(A) - \mu(A)),$$

where the class of sets $\sigma(\mathcal{P}_n)$ contains all sets obtained by unions of cells of $\mathcal{P}_n$. Therefore, for any $s > 0$, by the Markov inequality

$$\mathbb{P}\{L_n > \epsilon\} = \mathbb{P}\{L_n/2 > \epsilon/2\} = \mathbb{P}\{e^{nsL_n/2} > e^{ns\epsilon/2}\} \le \frac{\mathbb{E}\{e^{nsL_n/2}\}}{e^{ns\epsilon/2}}.$$

Moreover,

$$\begin{aligned}\mathbb{E}\{e^{snL_n/2}\} &= \mathbb{E}\{\max_{A\in\sigma(\mathcal{P}_n)} e^{sn(\mu_n(A)-\mu(A))}\} \\ &\leq \sum_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{sn(\mu_n(A)-\mu(A))}\} \\ &\leq 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{sn(\mu_n(A)-\mu(A))}\} \\ &= 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{sn\mu_n(A)}\} e^{-sn\mu(A)}.\end{aligned}$$

For any fixed Borel set A,

$$\mathbb{E}\{e^{sn\mu_n(A)}\} = \mathbb{E}\{e^{s\sum_{i=1}^n \mathbb{I}_{\mathbf{x}_i\in A}}\} = \prod_{i=1}^n \mathbb{E}\{e^{s\mathbb{I}_{\mathbf{x}_i\in A}}\} = \left(e^s\mu(A) + 1 - \mu(A)\right)^n.$$

Thus, for any $s > 0$, we have that

$$\mathbb{P}\{L_n > \epsilon\} \leq 2^{m_n}\left[\max_{A\in\sigma(\mathcal{P}_n)} e^{-s(\mu(A)+\epsilon/2)}\left(e^s\mu(A) + 1 - \mu(A)\right)\right]^n.$$

For fixed set A, choose

$$e^s = \frac{\mu(A)+\epsilon/2}{1-(\mu(A)+\epsilon/2)}\frac{1-\mu(A)}{\mu(A)},$$

then for this s,

$$\begin{aligned}e^{-s(\mu(A)+\epsilon/2)}\left(e^s\mu(A) + 1 - \mu(A)\right) &= e^{-I((\mu(A)+\epsilon/2,1-\mu(A)-\epsilon/2),(\mu(A),1-\mu(A)))} \\ &\leq e^{-\epsilon^2/2},\end{aligned}$$

where the last step follows from the Pinsker inequality. Thus,

$$\mathbb{P}\{L_n > \epsilon\} \leq 2^{m_n} e^{-n\epsilon^2/2}.$$

□

4.3 L_1-distance-based strongly consistent test

Theorem 4.3 results in a strongly consistent test such that reject the null-hypothesis $\mathcal{H}_0$ if

$$L_n > c_1\sqrt{\frac{m_n}{n}},$$

where

$$c_1 > \sqrt{2 \ln 2} \approx 1.177.$$

Moreover, assume that the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \dots$ is asymptotically fine. (Cf. (2.5)). Then, under the null hypothesis $\mathcal{H}_0 = \{\nu = \mu\}$, the inequality in Theorem 4.3 implies an upper bound on the error of the first kind

$$\mathbb{P}\left\{ L_n > c_1 \sqrt{\frac{m_n}{n}} \right\} \le 2^{m_n} e^{-n c_1^2 m_n/(2n)} = e^{-m_n (c_1^2/2 - \ln 2)} \to 0$$

If $m_n / \ln n \to \infty$ then

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ L_n > c_1 \sqrt{\frac{m_n}{n}} \right\} < \infty,$$

therefore the Borel-Cantelli lemma implies that the goodness of fit test based on the statistic L_n is strongly consistent under the null hypothesis $\mathcal{H}_0$, independently of the underlying distribution μ.

Under the alternative hypothesis $\mathcal{H}_1 = \{\nu \neq \mu\}$, the triangle inequality implies that

$$\begin{aligned} L_n &= \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \mu(A_{nj})| \\ &\ge \sum_{j=1}^{m_n} |\mu(A_{nj}) - \nu(A_{nj})| - \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \nu(A_{nj})|. \end{aligned}$$

Because of the argument above,

$$\sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \nu(A_{nj})| \to 0,$$

a.s., while the condition (2.5) and $\{\nu \neq \mu\}$ imply that

$$\sum_{j=1}^{m_n} |\mu(A_{nj}) - \nu(A_{nj})| \to 2 \sup_B |\mu(B) - \nu(B)| = 2V(\mu, \nu) > 0. \tag{4.6}$$

therefore

$$\liminf_{n \to \infty} L_n \ge 2V(\mu, \nu) > 0 \tag{4.7}$$

a.s., therefore $L_n > c_1 \sqrt{m_n/n}$ a.s. for n large enough, and so the goodness of fit test based on L_n is strongly consistent under the alternative hypothesis $\mathcal{H}_1$, too.

In order to show (4.6) we apply the technique from Barron, Györfi and van der Meulen (1992). Choose a measure λ which dominates μ and ν, for example, $\lambda = \mu + \nu$, and denote by f the Radon-Nikodym derivative of $\mu - \nu$ with respect to λ. Then, on the one hand,

$$\begin{aligned}
\sum_{A\in\mathcal{P}_n} |\mu(A) - \nu(A)| &= \sum_{A\in\mathcal{P}_n} \left| \int_A f \, \mathrm{d}\lambda \right| \\
&\leq \sum_{A\in\mathcal{P}_n} \int_A |f| \, \mathrm{d}\lambda \\
&= \int |f| \, \mathrm{d}\lambda \\
&= 2 \sup_B |\mu(B) - \nu(B)|.
\end{aligned}$$

On the other hand, for uniformly continuous f, using (2.5),

$$\sum_{A\in\mathcal{P}_n} \left| \int_A f \, \mathrm{d}\lambda \right| \to \int |f| \, \mathrm{d}\lambda.$$

If f is arbitrary then, for a given $\delta > 0$, choose a uniformly continuous $\tilde{f}$ such that

$$\int |f - \tilde{f}| \, \mathrm{d}\lambda < \delta.$$

Thus

$$\begin{aligned}
\sum_{A\in\mathcal{P}_n} \left| \int_A f \, \mathrm{d}\lambda \right| &\geq \sum_{A\in\mathcal{P}_n} \left| \int_A \tilde{f} \, \mathrm{d}\lambda \right| - \sum_{A\in\mathcal{P}_n} \left| \int_A (f - \tilde{f}) \, \mathrm{d}\lambda \right| \\
&\geq \sum_{A\in\mathcal{P}_n} \left| \int_A \tilde{f} \, \mathrm{d}\lambda \right| - \int |f - \tilde{f}| \, \mathrm{d}\lambda \\
&\geq \sum_{A\in\mathcal{P}_n} \left| \int_A \tilde{f} \, \mathrm{d}\lambda \right| - \delta \\
&\to \int |\tilde{f}| \, \mathrm{d}\lambda - \delta \\
&\geq \int |f| \, \mathrm{d}\lambda - 2\delta \\
&= 2 \sup_B |\mu(B) - \nu(B)| - 2\delta.
\end{aligned}$$

The result follows since δ was arbitrary.

4.4 L_1-distance-based α-level test

Beirlant, Györfi and Lugosi (1994) proved, under conditions

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0,$$

and

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0,$$

that

$$\sqrt{n}\left(L_n - \mathbb{E}\{L_n\}\right)/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\xrightarrow{\mathcal{D}}$ indicates convergence in distribution and $\sigma^2 = 1 - 2/\pi$.

Let $\alpha \in (0,1)$. Consider the test which rejects $\mathcal{H}_0$ when

$$L_n \;>\; c_2\sqrt{\frac{m_n}{n}} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) \approx c_2\sqrt{\frac{m_n}{n}},$$

where

$$c_2 = \sqrt{2/\pi} \approx 0.798.$$

Then the test is asymptotically an α-level test.

Comparing c_2 above with c_1 in the strong consistent test, both tests behave identically with respect to $\sqrt{m_n/n}$ for large enough n, but c_2 is smaller.

Under $\mathcal{H}_0$,

$$\mathbb{P}\{\sqrt{n}(L_n - \mathbb{E}\{L_n\})/\sigma \le x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the asymptotically α-level test rejects the null hypothesis if

$$L_n > \mathbb{E}\{L_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha).$$

Beirlant, Györfi and Lugosi (1994) proved an upper bound

$$\mathbb{E}\{L_n\} \le \sqrt{2/\pi}\sqrt{\frac{m_n}{n}}.$$

Chapter 5

Testing Homogeneity

5.1 The testing problem

Consider two mutually independent samples of $\mathbb{R}^d$-valued random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{X}'_1, \ldots, \mathbf{X}'_n$ with *i.i.d.* components distributed according to unknown probability measures μ and μ'. We are interested in testing the null hypothesis that the two samples are homogeneous, that is

$$\mathcal{H}_0 : \mu = \mu'.$$

Such tests have been extensively studied in the statistical literature for special parametrized models, *e.g.* for linear or loglinear models. For example, the analysis of variance provides standard tests of homogeneity when μ and μ' belong to a normal family on the line. For multinomial models these tests are discussed in common statistical textbooks, together with the related problem of testing independence in contingency tables. For testing homogeneity in more general parametric models, we refer the reader to the monograph of Greenwood and Nikulin (1996) and further references therein.

However, in many real life applications, the parametrized models are either unknown or too complicated for obtaining asymptotically α-level homogeneity tests by the classical methods. For $d = 1$, there are nonparametric procedures for testing homogeneity, for example, the Cramer-Mises, Kolmogorov-Smirnov, Wilcoxon tests. The problem of $d > 1$ is much more complicated, but nonparametric tests based on finite partitions of $\mathbb{R}^d$ may provide a welcome alternative. Such results are the extensions of Read and Cressie (1988).

In the present chapter, we discuss a simple approach based on a L_1 distance test statistic. The advantage of our test procedure is that, besides being explicit and rela-

tively easy to carry out, it requires very few assumptions on the partition sequence, and it is consistent. Let us now describe our test statistic.

Denote by μ_n and μ'_n the empirical measures associated with the samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{X}'_1, \ldots, \mathbf{X}'_n$, respectively, so that

$$\mu_n(A) = \frac{\#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}}{n},$$

and, similarly,

$$\mu'_n(A) = \frac{\#\{i : \mathbf{X}'_i \in A, i = 1, \ldots, n\}}{n}.$$

Based on a finite partition $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ of $\mathbb{R}^d$ ($m_n \in \mathbb{N}^*$), we let the test statistic comparing μ_n and μ'_n be defined as

$$T_n = \sum_{j=1}^{m_n} |\mu_n(A_{n,j}) - \mu'_n(A_{n,j})|.$$

5.2 L_1-distance-based strongly consistent test

The following theorem extends the results of Beirlant, Devroye, Györfi and Vajda (2001), and Devroye and Györfi (2002) to the statistic T_n.

Theorem 5.1. (Biau, Györfi (2005).) *Assume that conditions*

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{5.1}$$

and

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0, \tag{5.2}$$

are satisfied. Then, under $\mathcal{H}_0$, for all $0 < \varepsilon < 2$,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\{T_n > \varepsilon\} = -g_T(\varepsilon),$$

where

$$g_T(\varepsilon) = (1 + \varepsilon/2) \ln(1 + \varepsilon/2) + (1 - \varepsilon/2) \ln(1 - \varepsilon/2).$$

PROOF. We prove only the upper bound

$$\mathbb{P}\{T_n > \epsilon\} \leq 2^{m_n} e^{-n g_T(\epsilon)} \leq 2^{m_n} e^{-n\epsilon^2/4}. \tag{5.3}$$

For any $s > 0$, the Markov inequality implies that

$$\mathbb{P}\{T_n > \epsilon\} \;=\; \mathbb{P}\{e^{snT_n} > e^{sn\epsilon}\} \leq \frac{\mathbb{E}\{e^{snT_n}\}}{e^{sn\epsilon}}.$$

By Scheffé's theorem for partitions

$$T_n = \sum_{A\in\mathcal{P}_n} |\mu_n(A) - \mu'_n(A)| = 2 \max_{A\in\sigma(\mathcal{P}_n)} \left(\mu_n(A) - \mu'_n(A)\right),$$

where the class of sets $\sigma(\mathcal{P}_n)$ contains all sets obtained by unions of cells of $\mathcal{P}_n$. Therefore

$$\begin{aligned}
\mathbb{E}\{e^{snT_n}\} &= \mathbb{E}\{\max_{A\in\sigma(\mathcal{P}_n)} e^{2sn(\mu_n(A)-\mu'_n(A))}\} \\
&\leq \sum_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{2sn(\mu_n(A)-\mu'_n(A))}\} \\
&\leq 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{2sn(\mu_n(A)-\mu'_n(A))}\} \\
&= 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{2sn\mu_n(A)}\}\mathbb{E}\{e^{-2sn\mu'_n(A)}\}.
\end{aligned}$$

Clearly,

$$\begin{aligned}
\mathbb{E}\{e^{2sn\mu_n(A)}\} &= \sum_{k=0}^{n} e^{2sk} \binom{n}{k} \mu(A)^k \left(1-\mu(A)\right)^{n-k} \\
&= \left(e^{2s}\mu(A) + 1 - \mu(A)\right)^n,
\end{aligned}$$

and, similarly, under $\mathcal{H}_0$,

$$\begin{aligned}
\mathbb{E}\{e^{-2sn\mu'_n(A)}\} &= \sum_{k=0}^{n} e^{-2sk} \binom{n}{k} \mu(A)^k \left(1-\mu(A)\right)^{n-k} \\
&= \left(e^{-2s}\mu(A) + 1 - \mu(A)\right)^n.
\end{aligned}$$

The remainder of the proof is under the null hypothesis $\mathcal{H}_0$. From above, we deduce that

$$\begin{aligned}
&\mathbb{E}\{e^{snT_n}\} \\
&\quad \leq 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \left(e^{2s}\mu(A) + 1 - \mu(A)\right)^n \left(e^{-2s}\mu(A) + 1 - \mu(A)\right)^n \\
&\quad = 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \left[\left(e^{2s}\mu(A) + 1 - \mu(A)\right)\left(e^{-2s}\mu(A) + 1 - \mu(A)\right)\right]^n \\
&\quad = 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \left[1 + \mu(A)\left(1 - \mu(A)\right)\left(e^{2s} + e^{-2s} - 2\right)\right]^n \\
&\quad \leq 2^{m_n} \left[1 + (e^{2s} + e^{-2s} - 2)/4\right]^n \\
&\quad = 2^{m_n} \left[1/2 + (e^{2s} + e^{-2s})/4\right]^n .
\end{aligned}$$

It implies that

$$\mathbb{P}\{T_n > \epsilon\} \quad \leq \quad \inf_{s>0} \frac{\mathbb{E}\{e^{snT_n}\}}{e^{sn\epsilon}} \leq 2^{m_n} \left[\inf_{s>0} \frac{1/2 + (e^{2s} + e^{-2s})/4}{e^{s\epsilon}}\right]^n$$

One can verify that the infimum is achieved at

$$e^{2s} = \frac{1 + \varepsilon/2}{1 - \varepsilon/2},$$

and then

$$\mathbb{P}\{T_n > \epsilon\} \leq 2^{m_n} e^{-n g_T(\epsilon)}.$$

The Pinsker inequality implies that

$$g_T(\epsilon) \geq \epsilon^2/4$$

therefore

$$\mathbb{P}\{T_n > \epsilon\} \leq 2^{m_n} e^{-n\epsilon^2/4}.$$

□

The technique of Theorem 5.1 yields a distribution-free strong consistent test of homogeneity, which rejects the null hypothesis if T_n becomes large. We insist on the fact that the test presented in Corollary 5.1 is entirely distribution-free, i.e., the measures μ and μ' are completely arbitrary.

Corollary 5.1. (Biau, Györfi (2005).) *Consider the test which rejects $\mathcal{H}_0$ when*

$$T_n > c_1 \sqrt{\frac{m_n}{n}},$$

where

$$c_1 > 2\sqrt{\ln 2} \approx 1.6651.$$

Assume that condition (5.1) is satisfied and

$$\lim_{n\to\infty} \frac{m_n}{\ln n} = \infty.$$

Then, under $\mathcal{H}_0$, after a random sample size the test makes a.s. no error. Moreover, if

$$\mu \neq \mu',$$

and the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is asymptotically fine, (cf. (2.5)), then after a random sample size the test makes a.s. no error.

PROOF. Under $\mathcal{H}_0$, by (5.3),

$$\begin{aligned}
\mathbb{P}\left\{T_n > c_1\sqrt{\frac{m_n}{n}}\right\} &\leq 2^{m_n} e^{-ng_T\left(c_1\sqrt{m_n/n}\right)} \\
&= 2^{m_n} e^{-nc_1^2(m_n/n)/4 + n\cdot o(m_n/n)} \\
&= e^{-\left(c_1^2/4 - \ln 2 + o(1)\right)m_n},
\end{aligned}$$

as $n \to \infty$. Therefore the condition $m_n/\ln n \to \infty$ implies that

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{T_n > c_1\sqrt{\frac{m_n}{n}}\right\} < \infty,$$

and by the Borel-Cantelli lemma we are ready with the first half of the corollary. Concerning the second half, in the same way as for (4.6) we can show that by the additional condition (2.5),

$$\liminf_{n\to\infty} T_n \geq 2 \sup_B |\mu(B) - \mu'(B)| > 0 \tag{5.4}$$

a.s. □

5.3 L_1-distance-based α-level test

Again, one can prove the following asymptotic normality:

Theorem 5.2. (Biau, Györfi (2005).) *Assume that conditions (5.1) and (5.2) are satisfied. Then, under $\mathcal{H}_0$, there exists a centering sequence $C_n = \mathbb{E}\{T_n\}$ such that*

$$\sqrt{n}\,(T_n - C_n)\,/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\sigma^2 = 2(1-2/\pi)$.

Theorem 5.2 yields the asymptotic null distribution of a consistent homogeneity test, which rejects the null hypothesis if T_n becomes large. In contrast to Corollary 5.1, and because of condition (5.2), this new test is *not* distribution-free. In particular, the measures μ and μ' have to be nonatomic.

Corollary 5.2. (Biau, Györfi (2005).) *Put $\alpha \in (0,1)$, and let $C^* \approx 0.7655$ denote a universal constant. Consider the test which rejects $\mathcal{H}_0$ when*

$$T_n > c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) \approx c_2\sqrt{\frac{m_n}{n}},$$

where

$$\sigma^2 = 2(1-2/\pi) \quad \textit{and} \quad c_2 = \frac{2}{\sqrt{\pi}} \approx 1.1284.$$

Then, under the conditions of Theorem 5.2, the test is an asymptotically α-level test. Moreover, under the additional condition (2.5), the test is consistent.

Proof. According to Theorem 5.2, under $\mathcal{H}_0$,

$$\mathbb{P}\{\sqrt{n}(T_n - \mathbb{E}\{T_n\})/\sigma \le x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the α-level test rejects the null hypothesis if

$$T_n > \mathbb{E}\{T_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha).$$

However, $\mathbb{E}\{T_n\}$ depends on the unknown distribution, thus we apply an upper bound on $\mathbb{E}\{T_n\}$, and so decrease the error probability. The following inequality is valid:

$$\mathbb{E}\{T_n\} \;\le\; c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n},$$

(cf. Biau, Györfi (2005)). Thus

$$\begin{aligned}\alpha &\approx \mathbf{P}\left\{T_n > \mathbb{E}\{T_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\} \\ &\geq \mathbf{P}\left\{T_n > c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\}.\end{aligned}$$

This proves that the test has asymptotic error probability at most α.
Under $\mu \neq \mu'$, the consistency of the test follows from (5.4). □

Chapter 6

Testing Independence

6.1 The testing problem

Consider a sample of $\mathbb{R}^d \times \mathbb{R}^{d'}$-valued random vectors $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ with independent and identically distributed (i.i.d.) pairs. The distribution of $(\mathbf{X}, \mathbf{Y})$ is denoted by ν, while μ_1 and μ_2 stand for the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. We are interested in testing the null hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ are independent,

$$\mathcal{H}_0 : \nu = \mu_1 \times \mu_2, \tag{6.1}$$

while making minimal assumptions regarding the distribution.

We obtain two kinds of tests for each statistic: first, we derive *strong consistent* tests — meaning that both on $\mathcal{H}_0$ and on its complement the tests make a.s. no error after a random sample size — based on large deviation bounds. While such tests are not common in the classical statistics literature, they are well suited to data analysis from streams, where we receive a sequence of observations rather than a sample of fixed size, and must return the best possible decision at each time using only current and past observations. Our strong consistent tests are *distribution-free*, meaning they require no conditions on the distribution being tested; and *universal*, meaning the test threshold holds independent of the distribution. Second, we obtain tests based on the asymptotic distribution of the L_1, which assume only that ν is nonatomic. Subject to this assumption, the tests are *consistent*: for a given asymptotic error rate on $\mathcal{H}_0$, the probability of error on $\mathcal{H}_1$ drops to zero as the sample size increases. Moreover, the thresholds for the asymptotic tests are distribution-independent. We emphasize that our tests are explicit, easy to carry out, and require very few assumptions on the partition sequences.

Additional independence testing approaches also exist in the statistics literature. For $d = d' = 1$, an early nonparametric test for independence, due to Hoeffding (1948),

Blum et al. (1961), De Wet (1980) is based on the notion of differences between the joint distribution function and the product of the marginals. The associated independence test is consistent under appropriate assumptions. Two difficulties arise when using this statistic in a test, however. First, quantiles of the null distribution are difficult to estimate. Second, and more importantly, the quality of the empirical distribution function estimates becomes poor as the dimensionality of the spaces $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ increases, which limits the utility of the statistic in a multivariate setting.

Rosenblatt (1975) defined the statistic as the L_2 distance between the joint density estimate and the product of marginal density estimates. Let K and K' be density functions (called kernels) defined on $\mathbb{R}^d$ and on $\mathbb{R}^{d'}$, respectively. For the bandwidth $h > 0$, define

$$K_h(\mathbf{x}) = \frac{1}{h^d} K\left(\frac{\mathbf{x}}{h}\right) \quad \text{and} \quad K'_h(\mathbf{y}) = \frac{1}{h^{d'}} K'\left(\frac{\mathbf{y}}{h}\right).$$

The Rosenblatt-Parzen kernel density estimates of the density of $(\mathbf{X}, \mathbf{Y})$ and $\mathbf{X}$ are respectively

$$f_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i) K'_h(\mathbf{y} - \mathbf{Y}_i) \text{ and } f_{n,1}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i), \tag{6.2}$$

with $f_{n,2}(\mathbf{y})$ defined by analogy. Rosenblatt (1975) introduced the kernel-based independence statistic

$$T_n = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} (f_n(\mathbf{x}, \mathbf{y}) - f_{n,1}(\mathbf{x}) f_{n,2}(\mathbf{y}))^2 d\mathbf{x}\, d\mathbf{y}. \tag{6.3}$$

Further approaches to independence testing can be employed when particular assumptions are made on the form of the distributions, for instance that they should exhibit symmetry. We do not address these approaches in the present study.

6.2 L_1-distance-based strongly consistent test

Denote by ν_n, $\mu_{n,1}$ and $\mu_{n,2}$ the empirical measures associated with the samples $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, respectively, so that

$$\begin{aligned} \nu_n(A \times B) &= n^{-1} \#\{i : (\mathbf{X}_i, \mathbf{Y}_i) \in A \times B, i = 1, \ldots, n\}, \\ \mu_{n,1}(A) &= n^{-1} \#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}, \quad \text{and} \\ \mu_{n,2}(B) &= n^{-1} \#\{i : \mathbf{Y}_i \in B, i = 1, \ldots, n\}. \end{aligned}$$

Given the finite partitions $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ of $\mathbb{R}^d$ and $\mathcal{Q}_n = \{B_{n,1}, \ldots, B_{n,m'_n}\}$ of $\mathbb{R}^{d'}$, we define the L_1 test statistic comparing ν_n and $\mu_{n,1} \times \mu_{n,2}$ as

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)|.$$

In the following two sections, we derive the large deviation and limit distribution properties of this L_1 statistic, and the associated independence tests.

For testing a simple hypothesis versus a composite alternative, Györfi and van der Meulen (1990) introduced a related goodness of fit test statistic L_n defined as

$$L_n(\mu_{n,1}, \mu_1) = \sum_{A \in \mathcal{P}_n} |\mu_{n,1}(A) - \mu_1(A)|.$$

Biau and Györfi (2005) proved that, for all $0 < \varepsilon$,

$$\mathbb{P}\{L_n(\mu_{n,1}, \mu_1) > \varepsilon\} \leq 2^{m_n} e^{-n\varepsilon^2/2}, \tag{6.4}$$

(cf. Theorem 4.3). We now describe a similar result for our L_1 independence statistic.

Theorem 6.1. *(*Gretton, Györfi (2010).*) Under $\mathcal{H}_0$, for all $0 < \varepsilon_1$, $0 < \varepsilon_2$ and $0 < \varepsilon_3$,*

$$\mathbb{P}\{L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \varepsilon_1 + \varepsilon_2 + \varepsilon_3\} \leq 2^{m_n \cdot m'_n} e^{-n\varepsilon_1^2/2} + 2^{m_n} e^{-n\varepsilon_2^2/2} + 2^{m'_n} e^{-n\varepsilon_3^2/2}.$$

Proof. We bound $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ according to

$$\begin{aligned}
L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) &= \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)| \\
&\leq \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu_n(A \times B) - \nu(A \times B)| \\
&\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu(A \times B) - \mu_1(A) \cdot \mu_2(B)| \\
&\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\mu_1(A) \cdot \mu_2(B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)|.
\end{aligned}$$

Under the null hypothesis $\mathcal{H}_0$, we have that

$$\sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu(A \times B) - \mu_1(A) \cdot \mu_2(B)| = 0.$$

Moreover

$$\begin{aligned}
& \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} |\mu_1(A)\cdot\mu_2(B) - \mu_{n,1}(A)\cdot\mu_{n,2}(B)| \\
\le\ & \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} |\mu_1(A)\cdot\mu_2(B) - \mu_1(A)\cdot\mu_{n,2}(B)| \\
& + \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} |\mu_1(A)\cdot\mu_{n,2}(B) - \mu_{n,1}(A)\cdot\mu_{n,2}(B)| \\
=\ & \sum_{B\in\mathcal{Q}_n} |\mu_2(B) - \mu_{n,2}(B)| + \sum_{A\in\mathcal{P}_n} |\mu_1(A) - \mu_{n,1}(A)| \\
=\ & L_n(\mu_{n,1},\mu_1) + L_n(\mu_{n,2},\mu_2).
\end{aligned}$$

Thus, (6.4) implies

$$\begin{aligned}
& \mathbb{P}\{L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) > \varepsilon_1+\varepsilon_2+\varepsilon_3\} \\
\le\ & \mathbb{P}\left\{L_n(\nu_n,\nu) > \varepsilon_1\right\} + \mathbb{P}\left\{L_n(\mu_{n,1},\mu_1) > \varepsilon_2\right\} + \mathbb{P}\left\{L_n(\mu_{n,2},\mu_2) > \varepsilon_3\right\} \\
\le\ & 2^{m_n\cdot m'_n} e^{-n\varepsilon_1^2/2} + 2^{m_n} e^{-n\varepsilon_2^2/2} + 2^{m'_n} e^{-n\varepsilon_3^2/2}.
\end{aligned}$$

□

Theorem 6.1 yields a strong consistent test of independence, which rejects the null hypothesis if $L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})$ becomes large. The test is distribution-free, i.e., the probability distributions ν, μ_1 and μ_2 are completely arbitrary; and the threshold is universal, i.e., it does not depend on the distribution.

Corollary 6.1. *(*Gretton, Györfi (2010).*) Consider the test which rejects $\mathcal{H}_0$ when*

$$L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) > c_1\left(\sqrt{\frac{m_n m'_n}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m'_n}{n}}\right) \approx c_1\sqrt{\frac{m_n m'_n}{n}},$$

where

$$c_1 > \sqrt{2\ln 2} \approx 1.177. \tag{6.5}$$

Assume that conditions

$$\lim_{n\to\infty}\frac{m_n m'_n}{n} = 0, \tag{6.6}$$

and

$$\lim_{n\to\infty}\frac{m_n}{\ln n} = \infty, \qquad \lim_{n\to\infty}\frac{m'_n}{\ln n} = \infty, \tag{6.7}$$

are satisfied. Then under $\mathcal{H}_0$, the test makes a.s. no error after a random sample size. Moreover, if

$$\nu \neq \mu_1 \times \mu_2,$$

and for any sphere S centered at the origin,

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n,\, A\cap S\neq 0} \operatorname{diam}(A) = 0 \tag{6.8}$$

and

$$\lim_{n\to\infty} \max_{B\in\mathcal{Q}_n,\, B\cap S\neq 0} \operatorname{diam}(B) = 0, \tag{6.9}$$

then after a random sample size the test makes a.s. no error.

PROOF. Under $\mathcal{H}_0$, we obtain from Theorem 6.1 a non-asymptotic bound for the tail of the distribution of $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$, namely

$$\begin{aligned}
& \mathbb{P}\left\{ L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) > c_1 \left(\sqrt{\frac{m_n m_n'}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m_n'}{n}} \right) \right\} \\
\leq\ & 2^{m_n m_n'} e^{-c_1^2 m_n m_n'/2} + 2^{m_n} e^{-c_1^2 m_n/2} + 2^{m_n'} e^{-c_1^2 m_n'/2} \\
\leq\ & e^{-(c_1^2/2 - \ln 2) m_n m_n'} + e^{-(c_1^2/2-\ln 2) m_n} + e^{-(c_1^2/2-\ln 2) m_n'}
\end{aligned}$$

as $n \to \infty$. Therefore the condition (6.7) implies

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) > c_1 \left(\sqrt{\frac{m_n m_n'}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m_n'}{n}} \right) \right\} < \infty,$$

and the proof under the null hypothesis is completed by the Borel-Cantelli lemma. For the result under the alternative hypothesis, we first apply the triangle inequality

$$\begin{aligned}
L_n(\nu_n, \mu_{n,1}\times\mu_{n,2}) \ \geq\ & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu(A\times B) - \mu_1(A)\cdot\mu_2(B)| \\
& - \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu_n(A\times B) - \nu(A\times B)| \\
& - \sum_{B\in\mathcal{Q}_n} |\mu_2(B) - \mu_{n,2}(B)| \\
& - \sum_{A\in\mathcal{P}_n} |\mu_1(A) - \mu_{n,1}(A)|.
\end{aligned}$$

The condition in (6.6) implies the three last terms of the right hand side tend to 0 a.s. Moreover, using the technique for (4.6) we can prove that by conditions (6.8) and (6.9),

$$\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}|\nu(A\times B)-\mu_1(A)\cdot\mu_2(B)|\to 2\sup_C|\nu(C)-\mu_1\times\mu_2(C)|>0$$

as $n\to\infty$, where the last supremum is taken over all Borel subsets C of $\mathbb{R}^d\times\mathbb{R}^{d'}$, and therefore

$$\liminf_{n\to\infty}L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\geq 2\sup_C|\nu(C)-\mu_1\times\mu_2(C)|>0 \tag{6.10}$$

a.s. □

6.3 L_1-distance-based α-level test

Again, one can prove the following asymptotic normality:

Theorem 6.2. *(*Gretton, Györfi (2010).*) Assume that conditions (6.6) and*

$$\lim_{n\to\infty}\max_{A\in\mathcal{P}_n}\mu_1(A)=0,\quad \lim_{n\to\infty}\max_{B\in\mathcal{Q}_n}\mu_2(B)=0, \tag{6.11}$$

are satisfied. Then, under $\mathcal{H}_0$, there exists a centering sequence $C_n=\mathbb{E}\{L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\}$ depending on ν such that

$$\sqrt{n}\left(L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-C_n\right)/\sigma\xrightarrow{\mathcal{D}}\mathcal{N}(0,1),$$

where $\sigma^2=1-2/\pi$.

Theorem 6.2 yields the asymptotic null distribution of a consistent independence test, which rejects the null hypothesis if $L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})$ becomes large. In contrast to Corollary 6.1, and because of condition (6.11), this new test is *not* distribution-free: the measures μ_1 and μ_2 have to be nonatomic.

Corollary 6.2. *(*Gretton, Györfi (2010).*) Let $\alpha\in(0,1)$. Consider the test which rejects $\mathcal{H}_0$ when*

$$\begin{aligned}L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) &> c_2\sqrt{\frac{m_nm_n'}{n}}+\frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\\ &\approx c_2\sqrt{\frac{m_nm_n'}{n}},\end{aligned}$$

where

$$\sigma^2 = 1 - 2/\pi \quad \textit{and} \quad c_2 = \sqrt{2/\pi} \approx 0.798.$$

Then, under the conditions of Theorem 6.2, the test is an asymptotically α-level test. Moreover, under the additional conditions (6.8) and (6.9), the test is consistent.

Before proceeding to the proof, we examine how the above test differs from that in Corollary 6.1. In particular, comparing c_2 above with c_1 in (6.5), both tests behave identically with respect to $\sqrt{m_n m'_n/n}$ for large enough n, but c_2 is smaller.

PROOF. According to Theorem 6.2, under $\mathcal{H}_0$,

$$\mathbb{P}\{\sqrt{n}(L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - C_n)/\sigma \le x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the α-level test rejects the null hypothesis if

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > C_n + \frac{\sigma}{\sqrt{n}}\, \Phi^{-1}(1-\alpha).$$

As C_n depends on the unknown distribution, we apply an upper bound

$$C_n = \mathbb{E}\{L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})\} \le \sqrt{2/\pi}\sqrt{\frac{m_n m'_n}{n}}$$

(cf. Gretton, Györfi (2010)). □

Bibliography

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6:127–132.

Arkadjew, A. G. and Braverman, E. M. (1966). *Zeichenerkennung und Maschinelles Lernen*. Oldenburg Verlag, München, Wien.

Audibert, J.-Y. and Tsybakov, A. B. (2005). Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35:608–633.

Barron, A. R. and Barron, R. L. (1988). Statistical learning networks: a unifying view. In *Proceedings of the 20-th Symposium on the Interface: Computing Science and Statistics*, Wegman, E. J., Gantz, D. T., and Miller, J. J., editors, pages 192–203. AMS, Alexandria, VA.

Barron, A. R., Györfi, L., and van der Meulen, E. C. (1992). Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, 38:1437–1454.

Barron, R. L. (1975). Learning networks improve computer-aided prediction and control. *Computer Design*, 75:65–70.

Beirlant, J., Devroye, L., Györfi, L., and Vajda, I. (2001). Large deviations of divergence measures on partitions. *Journal of Statistical Planning and Inference*, 93:1 – 16.

Beirlant, J., Györfi, L., and Lugosi, G. (1994). On the asymptotic normality of the l_1- and l_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318.

Bernstein, S. N. (1946). *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer-Verlag, New York.

Biau, G. and Györfi, L. (2005). On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51:3965–3973.

Biglieri, E. and Györfi, L. (2014). Some remarks on robust binary hypothesis testing. In *2014 IEEE International Symposium on Information Theory, Honolulu, Hawaii*, pages 566–570. IEEE.

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32:485–498.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, CA.

Broder, A. J. (1990). Strategies for efficient incremental nearest neighbor search. *Pattern Recognition*, 23:171–178.

Cacoullos, T. (1965). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18:179–190.

Chernoff, H. (1952). A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507.

Coomans, D. and Broeckaert, I. (1986). *Potential Pattern Recognition in Chemical and Medical Decision Making*. Research Studies Press, Letchworth, Hertfordshire, England.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1990). *Introduction to Algorithms*. MIT Press, Boston.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica*, 2:299–318.

Dasarathy, B. V. (1991). *Nearest Neighbor Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.

De Wet, T. (1980). Cramér-von Mises tests for independence. *Journal of Multivariate Analysis*, 10:38–50.

Devijver, P. A. (1980). An overview of asymptotic properties of nearest neighbor rules. In *Pattern Recognition in Practice*, Gelsema, E. S. and Kanal, L. N., editors, pages 343–350. Elsevier Science Publishers, Amsterdam.

Devroye, L. (1981a). On the asymptotic probability of error in nonparametric discrimination. *Annals of Statistics*, 9:1320–1327.

Devroye, L. (1981b). On the inequality of Cover and Hart in nearest neighbor discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:75–78.

Devroye, L. (1988). Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:530–543.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Devroye, L. and Györfi, L. (2002). Distribution and density estimation. In *Principles of Nonparametric Learning*, Györfi, L., editor, pages 223–286. Springer-Verlag, Wien.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Devroye, L., Györfi, L., and Lugosi, G. (2002). A note on robust hypothesis testing. *IEEE Transactions on Information Theory*, 48:2111–2114.

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23:71–82.

Devroye, L. and Wagner, T. J. (1976). Nonparametric discrimination and density estimation. Technical Report 183, Electronics Research Center, University of Texas.

Devroye, L. and Wagner, T. J. (1980a). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8:231–239.

Devroye, L. and Wagner, T. J. (1980b). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8:231–239.

Devroye, L. and Wagner, T. J. (1980c). On the L_1 convergence of kernel estimators of regression functions with applications in discrimination. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 51:15–21.

Devroye, L. and Wagner, T. J. (1982). Nearest neighbor methods in discrimination. In *Handbook of Statistics*, Krishnaiah, P. R. and Kanal, L., editors, volume 2, pages 193–197. North Holland, Amsterdam.

Döring, M. and L. Györfi, H. W. (2015). Exact rate of convergence of kernel-based classification rule. In *Challenges in Statistics and Data Mining*, S. Matwin, J. M., editor, pages 71–91. Springer-Verlag.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929.

Faragó, A., Linder, T., and Lugosi, G. (1993). Fast nearest neighbor search in dissimilarity spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:957–962.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications, Vol.1*. John Wiley, New York.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Fix, E. and Hodges, J. L. (1952). Discriminatory analysis: small sample performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Fix, E. and Hodges, J. L. (1991a). Discriminatory analysis, nonparametric discrimination, consistency properties. In *Nearest Neighbor Pattern Classification Techniques*, Dasarathy, B., editor, pages 32–39. IEEE Computer Society Press, Los Alamitos, CA.

Fix, E. and Hodges, J. L. (1991b). Discriminatory analysis: small sample performance. In *Nearest Neighbor Pattern Classification Techniques*, Dasarathy, B. V., editor, pages 40–56. IEEE Computer Society Press, Los Alamitos, CA.

Friedman, J. H. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, 26:404–408.

Friedman, J. H., Baskett, F., and Shustek, L. J. (1975). An algorithm for finding nearest neighbor. *IEEE Transactions on Computers*, 24:1000–1006.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3:209–226.

Fukunaga, K. and Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, 24:750–753.

Gadat, S., Klein, T., and Mateau, C. (2016). Classification with the nearest neighbor rule in general finite dimensional space. *Annals of Statistics*, 44:982–1009.

Greblicki, W. (1974). Asymptotically optimal probabilistic algorithms for pattern recognition and identification. Technical Report, Monografie No. 3, Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wroclawskiej No. 18, Wroclaw, Poland.

Greblicki, W. (1978a). Asymptotically optimal pattern recognition procedures with density estimates. *IEEE Transactions on Information Theory*, 24:250–251.

Greblicki, W. (1978b). Pattern recognition procedures with nonparametric density estimates. *IEEE Transactions on Systems, Man and Cybernetics*, 8:809–812.

Greblicki, W. (1981). Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities. *IEEE Transactions on Information Theory*, 27:364–366.

Greblicki, W. and Pawlak, M. (1981). Classification using the Fourier series estimate of multivariate density functions. *IEEE Transactions on Systems, Man and Cybernetics*, 11:726–730.

Greblicki, W. and Pawlak, M. (1982). A classification procedure using the multiple Fourier series. *Information Sciences*, 26:115–126.

Greblicki, W. and Pawlak, M. (1983). Almost sure convergence of classification procedures using Hermite series density estimates. *Pattern Recognition Letters*, 2:13–17.

Greblicki, W. and Pawlak, M. (1985). Pointwise consistency of the Hermite series density estimate. *Statistics and Probability Letters*, 3:65–69.

Greenwood, P. E. and Nikulin, M. S. (1996). *A Guide to Chi-Squared Testing*. Wiley, New York.

Gretton, A. and Györfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Multivariate Analysis*, 11:1391–1423.

Györfi, L. (1978). On the rate of convergence of nearest neighbor rules. *IEEE Transactions on Information Theory*, 29:509–512.

Györfi, L. and Györfi, Z. (1978). An upper bound on the asymptotic error probability of the k-nearest neighbor rule for multiple classes. *IEEE Transactions on Information Theory*, 24:512–514.

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

Györfi, L. and van der Meulen, E. C. (1990). A consistent goodness of fit test based on the total variation distance. In *Nonparametric Functional Estimation and Related Topics*, Roussas, G., editor, pages 631–645. Kluwer Academic Publishers, Dordrecht.

Györfi, L. and Walk, H. (2014). Strongly consistent detection for nonparametric hypotheses. In *Measures of Complexity: Festschrift in Honor of Alexey Chervonenkis*, Gammerman, A., Papadopoulos, H., and Vovk, V., editors. Springer, Heidelberg.

Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Annals of Statistics*, 36:2135–2152.

Hand, D. J. (1981). *Discrimination and Classification*. John Wiley, Chichester.

Haykin, S. (1992). *Adaptive Radar Detection and Estimation*. Wiley, New York.

Haykin, S. (1993). *Radar Array Processing*. Springer-Verlag, New York.

Helstrom, C. W. (1960). *Statistical Theory of Signal Detection*. Pergamon Press.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Hoeffding, W. (1948). A nonparametric test for independence. *The Annals of Mathematical Statistics*, 19:546–557.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Kang, E. W. (2008). *Radar System Analysis, Design, and Simulation*. Artech House, Boston, London.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*, pages 126–169. Springer Lecture Notes in Mathematics, Springer-Verlag, Berlin.

Kim, B. S. and Park, S. B. (1986). A fast k-nearest neighbor finding algorithm based on the ordered partition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:761–766.

Kohler, M. and Krzyżak, A. (2007). On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Transactions on Information Theory*, 53:1735–1742.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk USSR*, 114:953–956.

Krzyżak, A. (1986). The rates of convergence of kernel regression estimates and classification rules. *IEEE Transactions on Information Theory*, 32:668–679.

Krzyżak, A. and Pawlak, M. (1984). Distribution-free consistency of a nonparametric kernel regression estimate and classification. *IEEE Transactions on Information Theory*, 30:78–81.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127.

Levy, B. C. (2008). *Principles of Signal Detection and Parameter Estimation*. Springer, Berlin.

Lorentz, G. G. (1976). The thirteenth problem of Hilbert. In *Proceedings of Symposia in Pure Mathematics*, volume 28, pages 419–430. Providence, RI.

Lugosi, G. and Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistics*, 24:687–706.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829.

Meisel, W. (1969). Potential functions in mathematical pattern recognition. *IEEE Transactions on Computers*, 18:911–918.

Minsky, M. L. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, 15:134–137.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London A*, 231:289–337.

Niemann, H. and Goppert, R. (1988). An efficient branch-and-bound nearest neighbour classifier. *Pattern Recognition Letters*, 7:67–72.

Nilsson, N. J. (1965). *Learning Machines: Foundations of Trainable Pattern Classifying Systems*. McGraw-Hill, New York.

Papadimitriou, C. H. and Bentley, J. L. (1980). A worst-case analysis of nearest neighbor searching by projection. In *Automata, Languages and Programming 1980*, pages 470–482. Lecture Notes in Computer Science #85, Springer-Verlag, Berlin.

Papoulis, A. (1984). *Signal Analysis*. McGraw Hill, New York.

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. McGraw Hill, New York.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Read, T. and Cressie, N. (1988). *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis*. Springer-Verlag, New York.

Rejtő, L. and Révész, P. (1973). Density estimation and pattern classification. *Problems of Control and Information Theory*, 2:67–80.

Ripley, B. D. (1993). Statistical aspects of neural networks. In *Networks and Chaos—Statistical and Probabilistic Aspects*, Barndorff-Nielsen, O. E., Jensen, J. L., and Kendall, W. S., editors, pages 40–123. Chapman and Hall, London.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society*, 56:409–456.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, 3:1–14.

Ryzin, J. V. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya Series A*, 28:161–170.

Samworth, R. J. (2012). Optimal weighted nearest neighbor classifiers. *Annals of Statistics*, 40:2733–2763.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18:434–458.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Sebestyen, G. (1962). *Decision Making Processes in Pattern Recognition*. Macmillan, New York.

Sethi, I. K. (1981). A fast algorithm for recognizing nearest neighbors. *IEEE Transactions on Systems, Man and Cybernetics*, 11:245–248.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Skolnik, M. I. (1980). *Introduction to Radar Systems*. McGraw-Hill.

Specht, D. F. (1971). Series estimation of a probability density function. *Technometrics*, 13:409–424.

Steele, J. M. (1975). *Combinatorial entropy and uniform limit laws*. PhD Thesis, Stanford University, Stanford, CA.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer–Verlag, New York.

Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific, Singapore.

Trees, H. L. (1971). *Detection, Estimation, and Modulation Theory I.II.III.* Wiley, New York.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer–Verlag, New York.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280.

Vapnik, V. N. and Chervonenkis, A. Y. (1974a). Ordered risk minimization. I. *Automation and Remote Control*, 35:1226–1235.

Vapnik, V. N. and Chervonenkis, A. Y. (1974b). Ordered risk minimization. II. *Automation and Remote Control*, 35:1403–1412.

Vapnik, V. N. and Chervonenkis, A. Y. (1974c). *Theory of Pattern Recognition*. Nauka, Moscow. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

Vapnik, V. N. and Kotz, S. (2006). *Estimation of Dependences Based on Empirical Data*. Springer–Verlag, New York.

Vidal, E. (1986). An algorithm for finding nearest neighbors in (approximately) constant average time. *Pattern Recognition Letters*, 4:145–157.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, 26:359–372.

Weiss, S. M. and Kulikowski, C. A. (1991). *Computer Systems that Learn*. Morgan Kaufmann, San Mateo, CA.

Yunck, T. P. (1976). A technique to identify nearest neighbors. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:678–683.

NEMZETI KÖZSZOLGÁLATI EGYETEM