# The theory of statistical decisions

László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics
1521 Stoczek u. 2, Budapest, Hungary
gyorfi@cs.bme.hu

March 14, 2012

**Abstract**

In this note we survey the theory of statistical decisions, i.e., consider statistical inferences, where the target of the inference takes finitely many values. For the formulation of the Bayes decision, the aim is to minimize the weighted average of conditional error probabilities. In the scheme of simple statistical hypotheses testing we constrain a conditional error probability and minimize the other one. Study the composite hypotheses, the testing of homogeneity and the testing of independence, too. In the analysis the divergences ($L_1$-distance, I-divergence, Hellinger distance, etc.) between probability distributions play an important role.

# Contents

# 1 Bayes decision

## 1.1 Bayes risk

For the statistical inference, a $d$-dimensional observation vector $\mathbf{X}$ is given, and based on $\mathbf{X}$, the statistician has to make an inference on a random variable $Y$, which takes finitely many values, i.e., it takes values from the set $\{1, 2, \ldots, m\}$. 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

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

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 $g$,

$$\sum_{y=1}^{m} C(y, g^*(\mathbf{X})) P_y(\mathbf{X}) \leq \sum_{y=1}^{m} C(y, g(\mathbf{X})) P_y(\mathbf{X}). \tag{2}$$

**Theorem 1** *For any decision function $g$, we have that*

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

**Proof.** For a decision function $g$, let's calculate the risk.

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

where $\mathbb{I}$ denotes the indicator. (2) implies that

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

□

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

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

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

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

and the Bayes decision is of form

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

which is called maximum posteriori decision, too.

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

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

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

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

which has the form, for densities,

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

## 1.2 Approximation of Bayes decision

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 $R^*$.

**Lemma 1** *Put* $C_{max} = \max_{y,y'} C(y, y')$, *then*

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

**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

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

therefore

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

□

In the special case of the approximate maximum posteriori decision the inequality in Lemma 1 can be slightly improved:

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

Based on this relation, one can introduce efficient pattern recognition rules. (For the details, see Devroye, Györfi, and Lugosi [21].)

# 2 Testing simple hypotheses

## 2.1 $\alpha$-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 is 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 $0 < \alpha < 1$:

$$\min_{A:\, P_0(A^c) \leq \alpha} P_1(A). \tag{3}$$

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

**Theorem 2** (Neyman, Pearson [45]) *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 $A$*

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

*then*

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

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

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

The definition of $A_\gamma$ implies that

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

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

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

□

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

$$P_0(A^c_\gamma) = \alpha,$$

then the Neyman-Pearson Lemma implies that in order to solve the problem (3), it is enough to search for set of form $A_\gamma$, i.e.,

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

Then $A_\gamma$ is called the *most powerful $\alpha$-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 $\sigma^2$, 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 $\sigma^2$. 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

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

This 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 $\gamma'$, for which it is an $\alpha$-level test, i.e., the error of the first kind is $\alpha$:

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

Under the null hypothesis, the distribution of $\frac{1}{d}\sum_{i=1}^{d} X_i$ is normal with mean $m$ and with variance $\sigma^2/d$, 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 $\Phi$ denotes the standard normal distribution function, and so the critical value $\gamma'$ of an $\alpha$-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 $\sigma_0^2$. 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 $\alpha$-level test.

## 2.2 $\phi$-divergences

In the analysis of repeated observations the divergences between distribution play an important role. Imre Csiszár [14] introduced the concept of $\phi$-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 $\mu$ and $\nu$, let $\lambda$ be a $\sigma$-finite dominating measure of $\mu$ and $\nu$, for example, $\lambda = \mu + \nu$. Introduce the notations

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

and

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

Then the *$\phi$-divergence* of $\mu$ and $\nu$ is defined by

$$D_\phi(\mu, \nu) = \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}). \tag{4}$$

The Jensen inequality implies the most important property of the $\phi$-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, $\phi$ 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*
$$I(\mu,\nu) = D_{\phi_3}(\mu,\nu) = \int_{\mathbb{R}^d} \ln\left(\frac{g(\mathbf{x})}{f(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

- For
$$\phi_4(t) = (t-1)^2,$$
we get the $\chi^2$*-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 $\phi$-divergence is
$$D_\phi(\mu,\nu) = \sup_{\mathcal{P}} \sum_j \phi\left(\frac{\mu(A_j)}{\nu(A_j)}\right)\nu(A_j), \tag{5}$$
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
$$\begin{aligned}
D_\phi(\mu,\nu) &= \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \\
&= \sum_j \int_{A_j} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \\
&= \sum_j \frac{1}{\nu(A_j)} \int_{A_j} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x})\nu(A_j) \\
&\geq \sum_j \phi\left(\frac{1}{\nu(A_j)} \int_{A_j} \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\lambda(d\mathbf{x})\right)\nu(A_j) \\
&= \sum_j \phi\left(\frac{\mu(A_j)}{\nu(A_j)}\right)\nu(A_j).
\end{aligned}$$

The sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \dots$ 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 $\mathcal{P}_1, \mathcal{P}_2, \dots$ such that

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

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

$$\lim_{n \to \infty} \max_{A \in \mathcal{P}_n, A \cap S \neq 0} diam(A) = 0. \tag{6}$$

One can show that if the nested sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \dots$ 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 is verified in the particular case of $L_1$ distance. (Cf. Section 3.3.) 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 $\alpha_n$ and the error of the second kind $\beta_n$ tend to 0, if $n \to \infty$.

In order to formulate the Stein Lemma, we introduce the *I-divergence* (called also relative entropy)

$$D(f_0, f_1) = \int_{\mathbb{R}^d} f_0(\mathbf{x}) \ln \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} d\mathbf{x}, \tag{7}$$

(cf. Section 2.2).

The I-divergence is always non-negative:

$$-D(f_0, f_1) = \int_{\mathbb{R}^d} f_0(\mathbf{x}) \ln \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} d\mathbf{x} \leq \int_{\mathbb{R}^d} f_0(\mathbf{x}) \left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} - 1\right) d\mathbf{x} = 0.$$

**Theorem 3** (Stein [58]) *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

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

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 $\alpha_n$, 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 $\beta_n$ we have the following simple bound:

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

□

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 4** (Chernoff [12]). *Construct a test such that accept the null hypothesis $\mathcal{H}_0$ if*

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

*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

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

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

$$He(f_0, f_1) = \int_{\mathbb{R}^d} f_1(\mathbf{x})^{1/2} f_0(\mathbf{x})^{1/2} d\mathbf{x} \tag{8}$$

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 Section 2.2. One can check that

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

REMARK 3. Besides the concept of $\alpha$-level consistency, there is a new kind of consistency, called *strong consistency*, meaning that both on $\mathcal{H}_0$ and on its complement the tests make a.s. no error after a random sample size. In other words, denoting by $\mathbb{P}_0$ (*resp.* $\mathbb{P}_1$) the probability under the null hypothesis (*resp.* under the alternative), we have

$$\mathbb{P}_0\{\text{rejecting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1 \tag{9}$$

and

$$\mathbb{P}_1\{\text{accepting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1. \tag{10}$$

Because of the Chernoff bound, both errors tend to 0 exponentially fast, so the Borel-Cantelli Lemma implies that the maximum likelihood test is strongly consistent. In a real life problem, for example, when we get the data sequentially, one gets data just once, and should make good inference for these data. Strong consistency means that the single sequence of inference is a.s. perfect if the sample size is large enough. This concept is close to the definition of discernability introduced by Dembo and Peres [18]. For a discussion and references, we refer the reader to Devroye and Lugosi [23].

# 3 Testing simple versus composite hypotheses

## 3.1 Total variation and I-divergence

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

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

where the supremum is taken over all Borel sets $A$. The Scheffé Theorem below shows that the total variation is the half of the $L_1$ distance of the corresponding densities.

**Theorem 5** *(*SCHEFFÉ [55]*) If $\mu$ and $\nu$ are absolutely continuous 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*

$$L_1(f_0, f_1) = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} \tag{11}$$

*is called $L_1$-*distance*, cf. Section 2.2.)*

**Proof.** Note that

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

□

The Scheffé Theorem implies an equivalent definition of the total variation:

$$V(\mu,\nu)=\frac{1}{2}\sup_{\{A_j\}}\sum_j|\mu(A_j)-\nu(A_j)|, \tag{12}$$

where the supremum is taken over all finite Borel measurable partitions $\{A_j\}$.

The *information divergence* (also called I-divergence, Kullback-Leibler number, relative entropy) of $\mu$ and $\nu$ is defined by

$$I(\mu,\nu)=\sup_{\{A_j\}}\sum_j\mu(A_j)\ln\frac{\mu(A_j)}{\nu(A_j)}, \tag{13}$$

where the supremum is taken over all finite Borel measurable partitions $\{A_j\}$.

If the densities $f$ and $g$ exist then one can prove that

$$I(\mu,\nu)=D(f,g)=\int_{\mathbb{R}^d}f(\mathbf{x})\ln\frac{f(\mathbf{x})}{g(\mathbf{x})}d\mathbf{x}.$$

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

**Theorem 6** ( Csiszár [14], Kullback [39] and Kemperman [38])

$$2\{V(\mu,\nu)\}^2\le I(\mu,\nu). \tag{14}$$

**Proof.** Applying the notations of the proof of the Scheffé Theorem, put

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

then the Scheffé Theorem implies that

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

Moreover, from (13) we get that

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

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

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

which follows from the facts on the derivative:

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

□

## 3.2 Large deviation of $L_1$ distance

Consider the sample of $\mathbb{R}^d$-valued random vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ with *i.i.d.* components such that the common distribution is denoted by $\nu$. For a fixed distribution $\mu$, 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 $T_n = T_n(\mathbf{X}_1, \dots, \mathbf{X}_n)$.

For testing a simple hypothesis $\mathcal{H}_0$ that the distribution of the sample is $\mu$, versus a composite alternative, Györfi and van der Meulen [31] 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 $\mu_n$ denotes the empirical measures associated with the sample $\mathbf{X}_1, \dots, \mathbf{X}_n$, so that

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

for any Borel subset $A$, and $\mathcal{P}_n = \{A_{n,1}, \dots, A_{n,m_n}\}$ is a finite partition of $\mathbb{R}^d$. These authors also showed that under $\mathcal{H}_0$

$$\mathbb{P}(L_n \ge \epsilon) \le e^{-n(\frac{\epsilon^2}{8} + o(1))}.$$

Next we characterize the large deviation properties of $L_n$:

**Theorem 7** (Beirlant, Devroye, Györfi and Vajda [6]). *Assume that*

$$\lim_{n\to\infty} \max_j \mu(A_{n,j}) = 0 \tag{15}$$

*and*

$$\lim_{n\to\infty} \frac{m_n \ln n}{n} = 0. \tag{16}$$

*Then for all* $0 < \epsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\{L_n > \epsilon\} = -g_L(\epsilon), \tag{17}$$

*where*

$$g_L(\epsilon) = \inf_{0<p<1-\epsilon/2} \left( p \ln \frac{p}{p+\epsilon/2} + (1-p) \ln \frac{1-p}{1-p-\epsilon/2} \right). \tag{18}$$

Remark 4. Note that a lower bound for $g_L$ follows from Pinsker's inequality (14) such that

$$g_L(\epsilon) \ge \epsilon^2/2.$$

The best known lower bound is due to Toussaint [61]:

$$g_L(\epsilon) \ge \epsilon^2/2 + \epsilon^4/36 + \epsilon^6/280.$$

An upper bound $\hat{g}(\epsilon)$ of $g_L(\epsilon)$ can be obtained substituting $p$ by $\frac{1-\epsilon/2}{2}$ in definition of $g_L(\epsilon)$. Then

$$\hat{g}(\epsilon) = \frac{\epsilon}{2} \ln \frac{2+\epsilon}{2-\epsilon} \ge g_L(\epsilon)$$

(Vajda [66]). Further bounds can be found on p. 294-295 in Vajda [65]. Remark that also in Lemma 5.1 in Bahadur [2] it was observed that

$$g_L(\epsilon) = \frac{\epsilon^2}{2}(1 + o(1))$$

as $\epsilon \to 0$. The observations above mean that

$$\mathbb{P}\{L_n > \varepsilon\} \approx e^{-n g_L(\varepsilon)} \le e^{-n\varepsilon^2/2}.$$

In the proof of Theorem 7 we shall use the following lemma.

**Lemma 2** (Sanov [54], see p. 16 in Dembo, Zeitouni [19], or Problem 1.2.11 in Csiszár and Körner [15]). *Let $\Sigma$ be a finite set (alphabet), $\mathcal{L}_n$ be a set of types (possible empirical distributions) on $\Sigma$, and let $\Gamma$ be a set of distributions on $\Sigma$. If $Z_1, \ldots, Z_n$ are i.i.d. random variables taking values in $\Sigma$ and with distribution $\mu$ and $\mu_n$ denotes the empirical distribution then*

$$\left| \frac{1}{n} \ln \mathbb{P}\{\mu_n^* \in \Gamma\} + \inf_{\tau \in \Gamma \cap \mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{|\Sigma| \ln(n+1)}{n} \tag{19}$$

*where $|\Sigma|$ denotes the cardinality of $\Sigma$.*

**Proof.** Without loss of generality assume that $\Sigma = \{1, \ldots, m\}$. We shall prove that

$$\mathbb{P}\{\mu_n \in \Gamma\} \leq |\mathcal{L}_n| e^{-n \min_{\tau \in \Gamma} I(\tau, \mu)}$$

and

$$\mathbb{P}\{\mu_n \in \Gamma\} \geq \frac{1}{|\mathcal{L}_n|} e^{-n \min_{\tau \in \Gamma} I(\tau, \mu)}.$$

Because of our assumptions

$$\begin{aligned}
\mathbb{P}\{Z_1 = z_1, \ldots Z_n = z_n\} &= \prod_{i=1}^n \mathbb{P}\{Z_i = z_i\} \\
&= \prod_{i=1}^n \mu(z_i) \\
&= e^{\sum_{i=1}^n \ln \mu(z_i)} \\
&= e^{\sum_{i=1}^n \sum_{j=1}^m I_{z_i = j} \ln \mu(z_i)} \\
&= e^{\sum_{i=1}^n \sum_{j=1}^m I_{z_i = j} \ln \mu(j)} \\
&= e^{\sum_{j=1}^m n\mu_n(j) \ln \mu(j)} \\
&= e^{-n(H(\mu_n) + I(\mu_n, \mu))} \\
&=: P_\mu(z_1^n),
\end{aligned}$$

where $H(\mu_n)$ stands for the Shannon entropy for the distribution $\mu_n$. For any probability distribution $\tau \in \mathcal{L}_n$ we can define a probability distribution $P_\tau(z_1^n)$ in this way:

$$P_\tau(z_1^n) := e^{-n(H(\mu_n) + I(\mu_n, \tau))}.$$

Put

$$T_n(\tau) = \{z_1^n : \mu_n(z_1^n) = \tau\},$$

then

$$1 \geq P_\tau\{\mu_n = \tau\} = P_\tau\{z_1^n \in T_n(\tau)\} = |T_n(\tau)|e^{-nH(\tau)}$$

therefore

$$|T_n(\tau)| \leq e^{nH(\tau)},$$

which implies the upper bound:

$$\begin{aligned} \mathbb{P}\{\mu_n \in \Gamma\} &= \sum_{\tau\in\Gamma} P_\mu\{\mu_n = \tau\} \\ &\leq |\mathcal{L}_n| \max_{\tau\in\Gamma} P_\mu\{\mu_n = \tau\} \\ &= |\mathcal{L}_n| \max_{\tau\in\Gamma} |T_n(\tau)|e^{-n(H(\tau)+I(\tau,\mu))} \\ &\leq |\mathcal{L}_n| \max_{\tau\in\Gamma} e^{-nI(\tau,\mu)} \\ &= |\mathcal{L}_n|e^{-n\min_{\tau\in\Gamma} I(\tau,\mu)}. \end{aligned}$$

Concerning the lower bound notice that for any probability distribution $\nu \in \mathcal{L}_n$

$$\begin{aligned} \frac{P_\tau\{\mu_n = \tau\}}{P_\tau\{\mu_n = \nu\}} &= \frac{|T_n(\tau)|\prod_{a\in\Sigma}\tau(a)^{n\tau(a)}}{|T_n(\nu)|\prod_{a\in\Sigma}\tau(a)^{n\nu(a)}} \\ &= \prod_{a\in\Sigma} \frac{(n\nu(a))!}{(n\tau(a))!}\tau(a)^{n(\tau(a)-\nu(a))} \\ &\geq 1. \end{aligned}$$

This last inequality can be seen as follows: the terms of the last product are of the forms $\frac{m!}{l!}\left(\frac{l}{n}\right)^{l-m}$. It is easy to check that $\frac{m!}{l!} \geq l^{m-l}$, therefore

$$\prod_{a\in\Sigma} \frac{(n\nu(a))!}{(n\tau(a))!}\tau(a)^{n(\tau(a)-\nu(a))} \geq \prod_{a\in\Sigma} n^{n(\tau(a)-\nu(a))} = n^{n(\sum_{a\in\Sigma}\tau(a)-\sum_{a\in\Sigma}\nu(a))} = 1.$$

It implies that

$$P_\tau\{\mu_n = \tau\} \geq P_\tau\{\mu_n = \nu\}$$

and thus

$$\begin{aligned} 1 &= \sum_\nu P_\tau\{\mu_n = \nu\} \\ &\leq |\mathcal{L}_n|P_\tau\{\mu_n = \tau\} \\ &= |\mathcal{L}_n||T_n(\tau)|e^{-nH(\tau)}, \end{aligned}$$

consequently

$$|T_n(\tau)| \geq \frac{1}{|\mathcal{L}_n|} e^{nH(\tau)}.$$

This implies the lower bound:

$$\begin{aligned} \mathbb{P}\{\mu_n \in \Gamma\} &= \sum_{\tau\in\Gamma} P_\mu\{\mu_n = \tau\} \\ &\geq \max_{\tau\in\Gamma} P_\mu\{\mu_n = \tau\} \\ &= \max_{\tau\in\Gamma} |T_n(\tau)| e^{-n(H(\tau)+I(\tau,\mu))} \\ &\geq \frac{1}{|\mathcal{L}_n|} \max_{\tau\in\Gamma} e^{-nI(\tau,\mu)} \\ &= \frac{1}{|\mathcal{L}_n|} e^{-n \min_{\tau\in\Gamma} I(\tau,\mu)}. \end{aligned}$$

**Proof of Theorem 7.** Introduce the notation

$$D(\alpha\|\beta) = \alpha \ln \frac{\alpha}{\beta} + (1-\alpha) \ln \frac{1-\alpha}{1-\beta}. \tag{20}$$

Let $\bar{\mu}_n$ and $\mu_n^*$ denote the restrictions of $\mu$ and $\mu_n$ to the partition $\mathcal{P}_n$. We apply (19) for

$$\Sigma = \{A_{n,1}, \ldots, A_{n,m_n}\}$$

such that

$$\Gamma = \{\tau : 2V(\bar{\mu}_n, \tau) \geq \epsilon\}.$$

Then, according to (19),

$$\left| \frac{1}{n} \ln \mathbb{P}\{L_n \geq \epsilon\} + \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) \right| \leq \frac{m_n \ln(n+1)}{n}$$

and therefore, under (16),

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\{L_n > \epsilon\} = - \lim_{n\to\infty} \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n).$$

It now remains to show that

$$\lim_{n\to\infty} \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) = g(\epsilon).$$

The distributions in $\mathcal{L}_n$ are possible empirical distributions, having components of the form $\frac{r}{n}$, where $r$ is integer. Because of (15) we have that

$$m_n \to \infty,$$

therefore because of the continuity of $V(\tau, \bar{\mu}_n)$ and $I(\tau, \bar{\mu}_n)$

$$\lim_{n\to\infty} \inf_{\tau\in\Gamma\cap\mathcal{L}_n} I(\tau, \bar{\mu}_n) = \lim_{n\to\infty} \inf_{2V(\tau,\bar{\mu}_n)\geq\epsilon} I(\tau, \bar{\mu}_n).$$

Here

$$I(\tau, \bar{\mu}_n) = \sum_{j=1}^{m_n} \tau(A_{n,j}) \ln \frac{\tau(A_{n,j})}{\mu(A_{n,j})}.$$

Put

$$L = \{j : \mu(A_{n,j}) > \tau(A_{n,j})\}$$

and

$$A_n = \cup_{j\in L} A_{n,j}.$$

Then

$$2V(\tau, \bar{\mu}_n) = 2(\mu(A_n) - \tau(A_n))$$

and, by the Information Processing Theorem of Csiszár [14] (cf. the definition (13)),

$$I(\tau, \bar{\mu}_n) \geq D(\tau(A_n)\|\mu(A_n)),$$

where the equality holds iff $\frac{\tau(A_{n,j})}{\mu(A_{n,j})}$ is constant both on $L$ and $L^c$. Thus

$$\begin{aligned}
& \lim_{n\to\infty} \inf_{2V(\tau,\bar{\mu}_n)\geq\epsilon} I(\tau, \bar{\mu}_n) \\
= \; & \inf_{0<p<1-\epsilon/2:\tau(A_n)=p,\mu(A_n)=p+\epsilon/2} D(\tau(A_n)\|\mu(A_n)), \\
= \; & \inf_{0<p<1-\epsilon/2} \left( p \ln \frac{p}{p+\epsilon/2} + (1-p) \ln \frac{1-p}{1-p-\epsilon/2} \right) \\
= \; & g_L(\epsilon),
\end{aligned}$$

and Theorem 7 is proved. □

Biau and Györfi [10] provided an alternative derivation of $g_L(\varepsilon)$ and non-asymptotic upper bound.

**Theorem 8** (Biau and Györfi [10]). *For any $\epsilon > 0$,*

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

**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

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

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 $A$,

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

where $\mathbb{I}$ stands for the indicator. Thus, for any $s > 0$, we have that

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

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 $s$,

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

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

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

□

Remark 5. As a special case of relative frequencies, in the previous proof the Chernoff inequality

$$\mathbb{P}\{\mu_n(A) - \mu(A) \geq \epsilon\} \leq e^{-nD(\mu(A)+\epsilon\|\mu(A))}$$

and the Hoeffding inequality is contained:

$$\mathbb{P}\{\mu_n(A) - \mu(A) \geq \epsilon\} \leq e^{-2n\epsilon^2}. \tag{21}$$

The Hoeffding inequality can be extended as follows: 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 ($i = 1, \ldots, n$). Then, for all $\epsilon > 0$,

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

(Cf. Hoeffding [34].) A further refinement is the Berstein inequality such that it takes into account the variances, too: 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

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Var}\{X_i\} > 0.$$

Then, for all $\epsilon > 0$,

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

(Cf. Berstein [9].)

### 3.3 $L_1$-distance-based strongly consistent test

Theorem 8 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 $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is asymptotically fine. (Cf. (6)). Then, under the null hypothesis $\mathcal{H}_0 = \{\nu = \mu\}$, the inequality in Theorem 8 implies an upper bound on the error of the first kind

$$\mathbb{P}\left\{L_n > c_1\sqrt{\frac{m_n}{n}}\right\} \leq 2^{m_n} e^{-nc_1^2 m_n/(2n)} = e^{-m_n(c_1^2/2-\ln 2)} \to 0$$

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 $\mu$.

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

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

Because of the argument above,

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

a.s., while the condition (6) and $\{\nu \neq \mu\}$ imply that

$$\sum_{j=1}^{m_n} |\mu(A_{nj}) - \nu(A_{nj})| \to 2\sup_B |\mu(B) - \nu(B)| = 2V(\mu, \nu) > 0. \tag{22}$$

therefore

$$\liminf_{n\to\infty} L_n \geq 2V(\mu,\nu) > 0 \tag{23}$$

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 (22) we apply the technique from Barron, Györfi and van der Meulen [4]. Choose a measure $\lambda$ which dominates $\mu$ and $\nu$, for example, $\lambda = \mu + \nu$, and denote by $f$ the Radon-Nikodym derivative of $\mu - \nu$ with respect to $\lambda$. 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 (6),

$$\sum_{A\in\mathcal{P}_n} \left|\int_A f \,\mathrm{d}\lambda\right| \to \int |f| \,\mathrm{d}\lambda.$$

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 $\delta$ was arbitrary.

## 3.4 $L_1$-distance-based $\alpha$-level test

Beirlant, Györfi and Lugosi [7] 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

$$\sqrt{n}\,(L_n - \mathbb{E}\{L_n\})\,/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

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 $\alpha$-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 $\alpha$-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 [7] proved an upper bound

$$\mathbb{E}\{L_n\} \le \sqrt{2/\pi}\sqrt{\frac{m_n}{n}}.$$

## 3.5 I-divergence-based strongly consistent test

In the literature on goodness-of-fit testing the *I-divergence statistic*, *Kullback-Leibler divergence*, or *log-likelihood statistic*,

$$I_n = \sum_{j=1}^{m_n} \mu_n(A_{n,j}) \ln \frac{\mu_n(A_{n,j})}{\mu(A_{n,j})},$$

plays an important role. We refer to Tusnády [62] and Barron [3] who first discussed the exponential character of the tails of $I_n$. Kallenberg [37], and Quine and Robinson [50] proved that, for all $\epsilon > 0$,

$$\mathbb{P}\{I_n > \epsilon\} \le \binom{n+m_n-1}{m_n-1} e^{-n\epsilon} \le e^{m_n \ln(n+m_n) - n\epsilon}. \qquad (24)$$

Applying Sanov's Theorem, one can prove this bound similarly to that of Theorem 7.

A strongly consistent test can be introduced such that the test rejects the null hypothesis $\mathcal{H}_0$ if

$$I_n \ge \frac{m_n(\ln(n+m_n)+1)}{n}.$$

Under $\mathcal{H}_0$, we obtain a non-asymptotic bound for the tail of the distribution of $I_n$:

$$\mathbb{P}\left\{I_n > \frac{m_n(\ln(n+m_n)+1)}{n}\right\} \ \leq \ e^{m_n \ln(n+m_n) - n\frac{m_n(\ln(n+m_n)+1)}{n}} = e^{-m_n}.$$

Therefore condition $m_n/\ln n \to \infty$ implies

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{I_n > \frac{m_n(\ln(n+m_n)+1)}{n}\right\} < \infty,$$

and by the Borel-Cantelli lemma we have strong consistency under the null hypothesis.

Under the alternative hypothesis the proof of strong consistency follows from Pinsker's inequality:

$$L_n^2 \leq 2I_n.$$

Therefore (6) and (23) imply that

$$\liminf_{n\to\infty} 2I_n \ \geq \ \liminf_{n\to\infty} L_n^2 \geq 4\sup_C |\nu(C) - \mu(C)|^2 > 0$$

a.s., where the supremum is taken over all Borel subsets $C$ of $\mathbb{R}^d$. In fact, under conditions (6), and

$$I(\nu, \mu) < \infty,$$

one may get

$$\lim_{n\to\infty} I_n = I(\nu, \mu) > 0$$

a.s.

## 3.6 I-divergence-based $\alpha$-level test

Concerning the limit distribution, Inglot et al. [35], and Györfi and Vajda [30] proved that under the conditions in the previous subsection,

$$\frac{2nI_n - m_n}{\sqrt{2m_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

This implies that for any real valued $x$,

$$\mathbb{P}\left\{\frac{2nI_n - m_n}{\sqrt{2m_n}} \geq x\right\} \ \to \ 1 - \Phi(x),$$

which results in a test rejecting the null hypotheses $\mathcal{H}_0$ if

$$\frac{2nI_n - m_n}{\sqrt{2m_n}} \geq \Phi^{-1}(1-\alpha),$$

or equivalently

$$I_n \geq \frac{\Phi^{-1}(1-\alpha)\sqrt{m_n}}{\sqrt{2}n} + \frac{m_n}{2n} \approx \frac{m_n}{2n}.$$

Note that unlike the $L_1$ case, the ratio of the strong consistent threshold to the threshold of asymptotic $\alpha$-level test increases for increasing $n$.

# 4 Robust detection: testing composite versus composite hypotheses

A model of robust detection may be formulated as follows: let $f^{(1)}, \ldots, f^{(k)}$ be fixed densities on $\mathbb{R}^d$ which are the nominal densities under $k$ hypotheses. We observe i.i.d. random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ according to a common density $f$. Under the hypothesis $\mathcal{H}_j$ ($j = 1, \ldots, k$) the density $f$ is a distorted version of $f^{(j)}$. This notion may be formalized in various ways. In this section we assume that the true density $f$ lies within a certain total variation distance of the underlying nominal density. More precisely, we assume that there exists a positive number $\epsilon$ such that for some $j \in \{1, \ldots, k\}$

$$\|f - f^{(j)}\| \leq \Delta_j - \epsilon,$$

where $\Delta_j \overset{\text{def}}{=} (1/2)\min_{i \neq j} \|f^{(i)} - f^{(j)}\|$. Here $\|f - g\| = \int |f - g|$ denotes the $L_1$ distance between two densities. Recall that by Scheffé's theorem half of the $L_1$ distance equals the total variation distance:

$$\|f - g\| = 2 \sup_{A \subset \mathbb{R}^d} \left| \int_A f - \int_A g \right| = 2 \int_{\{\mathbf{x}: f(\mathbf{x}) > g(\mathbf{x})\}} (f(\mathbf{x}) - g(\mathbf{x})) d\mathbf{x} \ ,$$

where the supremum is taken over all Borel sets of $\mathbb{R}^d$. Thus, we formally define the $k$ hypotheses by

$$\mathcal{H}_j = \left\{ f : \|f - f^{(j)}\| \leq \Delta_j - \epsilon \right\}, \quad j = 1, \ldots, k \ .$$

Introduce the empirical measure

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

where $\mathbb{I}$ denotes the indicator function and $A$ is a Borel set. Let $\mathcal{A}$ denote the collection of $k(k-1)/2$ sets of the form

$$A_{i,j} = \left\{ \mathbf{x} : f^{(i)}(\mathbf{x}) > f^{(j)}(\mathbf{x}) \right\} \ , \quad 1 \le i < j \le k \ .$$

The proposed test is the following: accept hypothesis $\mathcal{H}_j$ if

$$\max_{A\in\mathcal{A}} \left| \int_A f^{(j)} - \mu_n(A) \right| = \min_{i=1,\ldots,k} \max_{A\in\mathcal{A}} \left| \int_A f^{(i)} - \mu_n(A) \right| \ .$$

(In case there are several indices achieving the minimum, choose the smallest one.) The main result of this section is the following:

**Theorem 9** (Devroye, Györfi, Lugosi [22].) *For any $f \in \bigcup_{j=1}^k \mathcal{H}_j$*

$$\mathbb{P}\{\mathit{error}\} \le 2k(k-1)^2 e^{-n\epsilon^2/2}.$$

**Proof.** Without loss of generality, assume that $f \in \mathcal{H}_1$. Observe that by Scheffé's theorem,

$$\begin{aligned}
2\max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(1)} \right| &\le \|f - f^{(1)}\| \\
&\le \Delta_1 - \epsilon \\
&\le \frac{1}{2}\|f^{(1)} - f^{(j)}\| - \epsilon \\
&= \max_{A\in\mathcal{A}} \left| \int_A f^{(1)} - \int_A f^{(j)} \right| - \epsilon \\
&\le \max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(1)} \right| + \max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(j)} \right| - \epsilon
\end{aligned}$$

by the triangle inequality. Rearranging the obtained inequality, we get that

$$\max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(1)} \right| \le \max_{A\in\mathcal{A}} \left| \int_A f - \int_A f^{(j)} \right| - \epsilon \ .$$

Therefore,

$$
\begin{aligned}
& \mathbb{P}\{\text{error}\} \\
= \; & \mathbb{P}\left\{\exists j > 1 : \max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right|\right\} \\
\leq \; & (k-1)\max_{j>1}\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right|\right\} \\
= \; & (k-1)\max_{j>1}\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right. \\
& \qquad \left. < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right\}.
\end{aligned}
$$

The inequality derived above implies that

$$
\begin{aligned}
& \mathbb{P}\{\text{error}\} \\
\leq \; & (k-1)\max_{j>1}\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(j)}\right| + \epsilon\right. \\
& \qquad \left. < \max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right\} \\
\leq \; & (k-1)\max_{j>1}\mathbb{P}\left\{\left|\max_{A\in\mathcal{A}}\left|\int_A f^{(j)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(j)}\right|\right| > \frac{\epsilon}{2}\right\} \\
& +(k-1)\mathbb{P}\left\{\left|\max_{A\in\mathcal{A}}\left|\int_A f^{(1)} - \mu_n(A)\right| - \max_{A\in\mathcal{A}}\left|\int_A f - \int_A f^{(1)}\right|\right| > \frac{\epsilon}{2}\right\} \\
\leq \; & 2(k-1)\mathbb{P}\left\{\max_{A\in\mathcal{A}}\left|\int_A f - \mu_n(A)\right| > \frac{\epsilon}{2}\right\} \\
& \qquad \text{(by a double application of the triangle inequality)} \\
\leq \; & 2(k-1)|\mathcal{A}|\max_{A\in\mathcal{A}}\mathbb{P}\left\{\left|\int_A f - \mu_n(A)\right| > \frac{\epsilon}{2}\right\} \\
\leq \; & 2k(k-1)^2 e^{-n\epsilon^2/2} ,
\end{aligned}
$$

where in the last step we used Hoeffding's inequality [34] (cf. (21)). □

# 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 defined on the same probability space and distributed according to unknown probability measures $\mu$ and $\mu'$. 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 $\mu$ and $\mu'$ 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 [25] and further references therein.

However, in many real life applications, the parametrized models are either unknown or too complicated for obtaining asymptotically $\alpha$-level homogeneity tests by the classical methods. As explained in Pardo, Pardo and Vajda [47], this is typically the case in electroencephalographic (EEG) and electrocardiographic (ECG) biosignal analysis, or in speech source characterization. In such situations parametric families cannot be adopted with confidence, nonparametric tests should be used. 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. In this context, Pardo, Pardo and Vajda [47] recently presented a partition-based generalized likelihood ratio test of homogeneity and derived its asymptotic distribution under the null hypothesis, enabling to control the asymptotic test size. The results of these authors extend former results of Read and Cressie [52], and Pardo, Pardo and Zografos [48] on disparity statistics.

In the present paper, 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 relatively 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 $\mu_n$ and $\mu'_n$ 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

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

for any Borel subset $A$, 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 $\mu_n$ and $\mu'_n$ 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 [6], and Devroye and Györfi [20] to the statistic $T_n$.

**Theorem 10** (Biau, Györfi [10].) *Assume that conditions*

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{25}$$

*and*

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0, \tag{26}$$

*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

$$\mathbb{P}\{T_n > \epsilon\} \leq 2^{m_n} e^{-n g_T(\epsilon)} \leq 2^{m_n} e^{-n\epsilon^2/4}.$$

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

$$\mathbb{P}\{T_n > \epsilon\} \ \ \leq \ \ \inf_{s>0} \frac{\mathbb{E}\{e^{snT_n}\}}{e^{sn\epsilon}} \leq 2^{m_n} \left[\inf_{s>0} \frac{1/2+(e^{2s}+e^{-2s})/4}{e^{s\epsilon}}\right]^n$$

One can verify that the infimum is achieved at

$$e^{2s} = \frac{1+\varepsilon/2}{1-\varepsilon/2},$$

and then

$$\mathbb{P}\{T_n > \epsilon\} \leq 2^{m_n} e^{-ng_T(\epsilon)}.$$

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 10 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 1 is entirely distribution-free, i.e., the measures $\mu$ and $\mu'$ are completely arbitrary.

**Corollary 1** (Biau, Györfi [10].) *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 (25) 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. (6)), then after a random sample size the test makes a.s. no error.*

**Proof.** Under $\mathcal{H}_0$, we easily obtain from the proof of Theorem 10 (cf. (**??**) and (**??**)) a non-asymptotic bound for the tail of the distribution of $T_n$, namely

$$\mathbb{P}\{T_n > \varepsilon\} \leq \inf_{s>0} \frac{\mathbb{E}\{e^{snT_n}\}}{e^{sn\varepsilon}} \leq 2^{m_n} e^{-ng_T(\varepsilon)} \leq 2^{m_n} e^{-n\varepsilon^2/4}. \tag{27}$$

Thus, by (**??**),

$$\begin{aligned}
\mathbb{P}\left\{T_n > c_1\sqrt{\frac{m_n}{n}}\right\} &\leq 2^{m_n} e^{-ng_T\left(c_1\sqrt{m_n/n}\right)} \\
&= 2^{m_n} e^{-nc_1^2(m_n/n)/4 + no(m_n/n)} \\
&= e^{-\left(c_1^2/4 - \ln 2 + o(1)\right)m_n},
\end{aligned}$$

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 in Section 3.3 we can show that by the additional condition (6),

$$\liminf_{n\to\infty} T_n \geq 2\sup_B |\mu(B) - \mu'(B)| > 0 \tag{28}$$

a.s. □

## 5.3 $L_1$-distance-based $\alpha$-level test

Similarly to Section 3.4, one can prove the following asymptotic normality:

**Theorem 11** (Biau, Györfi [10].) *Assume that conditions* (*25*) *and* (*26*) *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 11 yields the asymptotic null distribution of a consistent homogeneity test, which rejects the null hypothesis if $T_n$ becomes large. In contrast to Corollary 1, and because of condition (26), this new test is *not* distribution-free. In particular, the measures $\mu$ and $\mu'$ have to be nonatomic.

**Corollary 2** (Biau, Györfi [10].) *Put* $\alpha \in (0,1)$*, and let* $C^* \approx 0.7655$ *denote a universal constant. Consider the test which rejects* $\mathcal{H}_0$ *when*

$$T_n > c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha),$$

*where*

$$\sigma^2 = 2(1-2/\pi) \quad \textit{and} \quad c_2 = \frac{2}{\sqrt{\pi}} \approx 1.1284,$$

*and where* $\Phi$ *denotes the standard normal distribution function. Then, under the conditions of Theorem 11, the test has asymptotic significance level* $\alpha$*. Moreover, under the additional condition* (*6*)*, the test is consistent.*

**Proof.** According to Theorem 11, 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 $\alpha$-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:

$$\mathbb{E}\{T_n\} \quad \leq \quad c_2\sqrt{\frac{m_n}{n}} + C^* \frac{m_n}{n},$$

(cf. Biau, Györfi [10]). Thus

$$\begin{aligned} \alpha \quad &\approx \quad \mathbf{P}\left\{T_n > \mathbb{E}\{T_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\} \\ &\geq \quad \mathbf{P}\left\{T_n > c_2\sqrt{\frac{m_n}{n}} + C^* \frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\}. \end{aligned}$$

This proves that the test has asymptotic error probability at most $\alpha$. Under $\mu \neq \mu'$, the consistency of the test follows from (28). □

Note that, by condition (25),

$$c_2\sqrt{\frac{m_n}{n}} + C^* \frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) = c_2\sqrt{\frac{m_n}{n}}\,(1 + \mathrm{o}(1))\,,$$

therefore the order of the threshold does not depend on the level $\alpha$.

# 6 Testing independence

## 6.1 The testing problem

Consider a sample of $\Re^d \times \Re^{d'}$-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 defined on the same probability space. The distribution of $(\mathbf{X}, \mathbf{Y})$ is denoted by $\nu$, while $\mu_1$ and $\mu_2$ 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,

$$\mathcal{H}_0 : \nu = \mu_1 \times \mu_2, \tag{29}$$

while making minimal assumptions regarding the distribution.

We consider two main approaches to independence testing. The first is to partition the underlying space, and to evaluate the test statistic on the resulting discrete empirical measures. Consistency of the test must then be verified as the partition is refined for increasing sample size. Previous multivariate hypothesis tests in this framework, using the $L_1$ divergence measure, include homogeneity tests (to determine whether two random variables have the same distribution), by Biau and Györfi [10]; and goodness-of-fit tests (for whether a random variable has a particular distribution), by Györfi and van der Meulen [31], and Beirlant et al. [7]. The log-likelihood has also been employed on discretised spaces as a statistic for goodness-of-fit testing, by Györfi and Vajda [30]. We provide generalizations of both the $L_1$ and log-likelihood based tests to the problem of testing independence, representing to our knowledge the first application of these techniques to independence testing.

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 $L_1$ and log-likelihood statistics, which assume only that $\nu$ 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 [33], Blum et al. [11], De Wet [17] 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 $\Re^d$ and $\Re^{d'}$ increases, which limits the utility of the statistic in a multivariate setting.

Rosenblatt [53] 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 $\Re^d$ and on $\Re^{d'}$, 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

$$f_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{X}_i) K'_h(\mathbf{y}-\mathbf{Y}_i) \text{and} f_{n,1}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{X}_i), \quad (30)$$

with $f_{n,2}(\mathbf{y})$ defined by analogy. Rosenblatt [53] introduced the kernel-based independence statistic

$$T_n = \int_{\Re^d \times \Re^{d'}} (f_n(\mathbf{x}, \mathbf{y}) - f_{n,1}(\mathbf{x}) f_{n,2}(\mathbf{y}))^2 d\mathbf{x}\, d\mathbf{y}. \quad (31)$$

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 $L_1$-based strongly consistent test

Denote by $\nu_n$, $\mu_{n,1}$ 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

$$\begin{aligned} \nu_n(A \times B) &= n^{-1} \#\{i : (\mathbf{X}_i, \mathbf{Y}_i) \in A \times B, i = 1, \ldots, n\}, \\ \mu_{n,1}(A) &= n^{-1} \#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}, \quad \text{and} \\ \mu_{n,2}(B) &= n^{-1} \#\{i : \mathbf{Y}_i \in B, i = 1, \ldots, n\}, \end{aligned}$$

for any Borel subsets $A$ and $B$. 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 $\nu_n$ 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 [31] introduced a related goodness of fit test statistic $L_n$ defined as

$$L_n(\mu_{n,1}, \mu_1) = \sum_{A \in \mathcal{P}_n} |\mu_{n,1}(A) - \mu_1(A)|.$$

Beirlant et al. [6], and Biau and Györfi [10] proved that, for all $0 < \varepsilon$,

$$\mathbb{P}\{L_n(\mu_{n,1}, \mu_1) > \varepsilon\} \leq 2^{m_n} e^{-n\varepsilon^2/2}, \tag{32}$$

(cf. Theorem 8). We now describe a similar result for our $L_1$ independence statistic.

**Theorem 12** *(*Gretton, Györfi [26].*) 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, (32) implies

$$\begin{aligned}
& \mathbb{P}\{L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})>\varepsilon_1+\varepsilon_2+\varepsilon_3\} \\
\le & \mathbb{P}\{L_n(\nu_n,\nu)>\varepsilon_1\}+\mathbb{P}\{L_n(\mu_{n,1},\mu_1)>\varepsilon_2\}+\mathbb{P}\{L_n(\mu_{n,2},\mu_2)>\varepsilon_3\} \\
\le & 2^{m_n\cdot m_n'}e^{-n\varepsilon_1^2/2}+2^{m_n}e^{-n\varepsilon_2^2/2}+2^{m_n'}e^{-n\varepsilon_3^2/2}.
\end{aligned}$$

□

Theorem 12 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 $\nu$, $\mu_1$ and $\mu_2$ are completely arbitrary; and the threshold is universal, i.e., it does not depend on the distribution.

**Corollary 3** *(*Gretton, Györfi [26].*) Consider the test which rejects $\mathcal{H}_0$ when*

$$L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})>c_1\left(\sqrt{\frac{m_nm_n'}{n}}+\sqrt{\frac{m_n}{n}}+\sqrt{\frac{m_n'}{n}}\right)\approx c_1\sqrt{\frac{m_nm_n'}{n}},$$

*where*

$$c_1>\sqrt{2\ln 2}\approx 1.177. \tag{33}$$

*Assume that conditions*

$$\lim_{n\to\infty}\frac{m_nm_n'}{n}=0, \tag{34}$$

*and*

$$\lim_{n\to\infty}\frac{m_n}{\ln n}=\infty,\qquad \lim_{n\to\infty}\frac{m_n'}{\ln n}=\infty, \tag{35}$$

*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,*

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n,\, A\cap S\neq 0} \mathrm{diam}(A) = 0 \tag{36}$$

*and*

$$\lim_{n\to\infty} \max_{B\in\mathcal{Q}_n,\, B\cap S\neq 0} \mathrm{diam}(B) = 0, \tag{37}$$

*then after a random sample size the test makes a.s. no error.*

**Proof.** Under $\mathcal{H}_0$, we obtain from Theorem 12 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 (35) 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 (34) implies the three last terms of the right hand side tend to 0 a.s. Moreover, using the technique from Section 3.3 we can prove that by conditions (36) and (37),

$$\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}|\nu(A\times B)-\mu_1(A)\cdot\mu_2(B)|\to 2\sup_C|\nu(C)-\mu_1\times\mu_2(C)|>0$$

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

$$\liminf_{n\to\infty}L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\geq 2\sup_C|\nu(C)-\mu_1\times\mu_2(C)|>0 \tag{38}$$

a.s. □

## 6.3 $L_1$-based $\alpha$-level test

Similarly to Sections 3.4 and 5.3, one can prove the following asymptotic normality:

**Theorem 13** *(*Gretton, Györfi [26].) *Assume that conditions (34) and*

$$\lim_{n\to\infty}\max_{A\in\mathcal{P}_n}\mu_1(A)=0,\quad \lim_{n\to\infty}\max_{B\in\mathcal{Q}_n}\mu_2(B)=0, \tag{39}$$

*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 $\nu$ 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 13 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 3, and because of condition (39), this new test is *not* distribution-free: the measures $\mu_1$ and $\mu_2$ have to be nonatomic.

**Corollary 4** *(*Gretton, Györfi [26].) *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*

$$\sigma^2 = 1 - 2/\pi \quad \text{and} \quad c_2 = \sqrt{2/\pi} \approx 0.798,$$

*and $\Phi$ denotes the standard normal distribution function. Then, under the conditions of Theorem 13, the test has asymptotic significance level $\alpha$. Moreover, under the additional conditions (36) and (37), the test is consistent.*

Before proceeding to the proof, we examine how the above test differs from that in Corollary 3. In particular, comparing $c_2$ above with $c_1$ in (33), 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 13, 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 $\alpha$-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 [26]). □

## 6.4 I-divergence-based strongly consistent test

In the literature on goodness-of-fit testing the *I-divergence statistic*, *Kullback-Leibler divergence*, or *log-likelihood statistic*,

$$I_n(\mu_{n,1}, \mu_1) = \sum_{j=1}^{m_n} \mu_{n,1}(A_{n,j}) \log \frac{\mu_{n,1}(A_{n,j})}{\mu_1(A_{n,j})},$$

plays an important role. For testing independence, the corresponding log-likelihood test statistic is defined as

$$I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} \nu_n(A \times B) \log \frac{\nu_n(A \times B)}{\mu_{n,1}(A) \cdot \mu_{n,2}(B)}.$$

The large deviation and the limit distribution properties of $I_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ can be derived from the properties of

$$I_n(\nu_n, \nu) = \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} \nu_n(A \times B) \log \frac{\nu_n(A \times B)}{\nu(A \times B)}.$$

We have that under $\mathcal{H}_0$,

$$\begin{aligned}
& I_n(\nu_n, \nu) - I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) \\
= & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} \nu_n(A \times B) \log \frac{\nu_n(A \times B)}{\nu(A \times B)} \\
& - \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} \nu_n(A \times B) \log \frac{\nu_n(A \times B)}{\mu_{n,1}(A) \cdot \mu_{n,2}(B)} \\
= & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} \nu_n(A \times B) \log \frac{\mu_{n,1}(A) \cdot \mu_{n,2}(B)}{\nu(A \times B)} \\
= & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} \nu_n(A \times B) \log \frac{\mu_{n,1}(A) \cdot \mu_{n,2}(B)}{\mu_1(A) \cdot \mu_2(B)},
\end{aligned}$$

therefore

$$\begin{aligned}
& I_n(\nu_n, \nu) - I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) \\
= & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} \nu_n(A \times B) \left( \log \frac{\mu_{n,1}(A)}{\mu_1(A)} + \log \frac{\mu_{n,2}(B)}{\mu_2(B)} \right) \\
= & \sum_{A\in\mathcal{P}_n} \mu_{n,1}(A) \log \frac{\mu_{n,1}(A)}{\mu_1(A)} + \sum_{B\in\mathcal{Q}_n} \mu_{n,2}(B) \log \frac{\mu_{n,2}(B)}{\mu_2(B)} \\
= & I_n(\mu_{n,1}, \mu_1) + I_n(\mu_{n,1}, \mu_1) \\
\geq & 0.
\end{aligned}$$

A large deviation based test can be introduced such that the test rejects the independence if

$$I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) \geq \frac{m_n m'_n(\log(n + m_n m'_n) + 1)}{n}.$$

Under $\mathcal{H}_0$, (24) implies a non-asymptotic bound for the tail of the distribution of $I_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$:

$$\begin{aligned} & \mathbb{P}\left\{ I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \frac{m_n m'_n(\log(n + m_n m'_n) + 1)}{n} \right\} \\ \leq & \ \mathbb{P}\left\{ I_n(\nu_n, \nu) > \frac{m_n m'_n(\log(n + m_n m'_n) + 1)}{n} \right\} \\ \leq & \ e^{m_n m'_n \log(n+m_n m'_n) - n\frac{m_n m'_n(\log(n+m_n m'_n)+1)}{n}} \\ = & \ e^{-m_n m'_n}. \end{aligned}$$

Therefore condition (35) implies

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \frac{m_n m'_n(\log(n + m_n m'_n) + 1)}{n} \right\} < \infty,$$

and by the Borel-Cantelli lemma we have strong consistency under the null hypothesis.

Under the alternative hypothesis the proof of strong consistency follows from the Pinsker's inequality:

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})^2 \leq 2I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}). \tag{40}$$

Therefore,

$$\begin{aligned} \liminf_{n\to\infty} 2I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) & \geq (\liminf_{n\to\infty} L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}))^2 \\ & \geq 4 \sup_C |\nu(C) - \mu_1 \times \mu_2(C)|^2 > 0 \end{aligned}$$

a.s., where the supremum is taken over all Borel subsets $C$ of $\mathbb{R}^d \times \mathbb{R}^{d'}$.

## 6.5 I-divergence-based $\alpha$-level test

Concerning the limit distribution, Inglot et al. [35], and Györfi and Vajda [30] proved that under (25) and (26),

$$\frac{2nI_n(\mu_{n,1}, \mu_1) - m_n}{\sqrt{2m_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \tag{41}$$

This implies that for any real valued $x$, under the conditions (34) and (39),

$$\begin{aligned} \mathbb{P}\left\{ \frac{2nI_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - m_n m'_n}{\sqrt{2m_n m'_n}} \geq x \right\} & \leq \mathbb{P}\left\{ \frac{2nI_n(\nu_n, \nu) - m_n m'_n}{\sqrt{2m_n m'_n}} \geq x \right\} \\ & \to 1 - \Phi(x), \end{aligned}$$

which results in a test rejecting the independence if

$$\frac{2nI_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - m_n m_n'}{\sqrt{2m_n m_n'}} \geq \Phi^{-1}(1-\alpha),$$

or equivalently

$$I_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) \geq \frac{\Phi^{-1}(1-\alpha)\sqrt{2m_n m_n'} + m_n m_n'}{2n}.$$

Note that unlike the $L_1$ case, the ratio of the strong consistent threshold to the asymptotic threshold increases for increasing $n$.

# References

[1] M. S. Ali, and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser B*, 28:131–140, 1966.

[2] R. R. Bahadur. *Some Limit Theorems in Statistics.* SIAM, Philadelphia, 1971.

[3] A. R. Barron. Uniformly powerful goodness of fit tests. *The Annals of Statistics*, 17:107–124, 1989.

[4] A. R. Barron, L. Györfi, and E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, 38:1437–1454, 1992.

[5] M. S. Bartlett. The characteristic function of a conditional statistic. *Journal of the London Mathematical Society*, 13:62–67, 1938.

[6] J. Beirlant, L. Devroye, L. Györfi, and I. Vajda. Large deviations of divergence measures on partitions. *Journal of Statistical Planning and Inference*, 93:1–16, 2001.

[7] J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the $l_1$- and $l_2$-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318, 1994.

[8] J. Beirlant and D. M. Mason. On the asymptotic normality of $l_p$-norms of empirical functionals. *Mathematical Methods of Statistics*, 4:1–19, 1995.

[9] S. N. Berstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

[10] G. Biau and L. Györfi. On the asymptotic properties of a nonparametric $l_1$-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51:3965–3973, 2005.

[11] J. R. Blum, J. Kiefer, and M. Rosenblatt. Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32:485–498, 1961.

[12] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Ann. Math. Stat.*, 23:493–507, 1952.

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

[14] I. Csiszár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[15] I. Csiszár, and J. Körner. *Information Theory: Coding Theorems for Memoryless Systems.* Academic Press, New York, 1981.

[16] T. Cover, and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, 1991.

[17] T. De Wet. Cramér-von Mises tests for independence. *Journal of Multivariate Analysis*, 10(1):38–50, 1980.

[18] A. Dembo and Y. Peres. A topological criterion for hypothesis testing. *The Annals of Statistics*, 22:106–117, 1994.

[19] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer-Verlag, New York, second edition, 1998.

[20] L. Devroye and L. Györfi. Distribution and density estimation. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 223–286, Wien, 2002. Springer-Verlag.

[21] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[22] L. Devroye, L. Györfi, G. Lugosi. A note on robust hypothesis testing", *IEEE Trans. Information Theory*, 48, pp. 2111-2114, 2002.

[23] L. Devroye and G. Lugosi. Almost sure classification of densities. *J. Nonparametr. Stat.*, 14:675–698, 2002.

[24] Giné, E., Mason, D. M. and Zaitsev, A. Yu (2003). The $L_1$-norm density estimator process, *Ann. Probab.*, 31, pp. 719-768.

[25] P. E. Greenwood and M. S. Nikulin. *A Guide to Chi-Squared Testing.* Wiley, New York, 1996.

[26] A. Gretton and L. Györfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

[27] P. Groeneboom, and G. R. Shorack. Large deviations of goodness of fit statistics and linear combinations of order statistics. *Ann. Probability*, 9:971–987, 1981.

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

[29] L. Györfi, F. Liese, I. Vajda, and E. C. van der Meulen. Distribution estimates consistent in $\chi^2$-divergence. *Statistics*, 32:31–57, 1998.

[30] L. Györfi and I. Vajda. Asymptotic distributions for goodness of fit statistics in a sequence of multinomial models. *Statistics and Probability Letters*, 56:57–67, 2002.

[31] L. Györfi and E. C. van der Meulen. A consistent goodness of fit test based on the total variation distance. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 631–645. Kluwer, Dordrecht, 1990.

[32] P. Hall. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, 14:1–16, 1984.

[33] W. Hoeffding. A nonparametric test for independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

[34] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[35] T. Inglot, T. Jurlewitz, and T. Ledwina. Asymptotics for multinomial goodness of fit tests for a simple hypothesis. *Theory of Probability and Its Applications*, 35:797–803, 1990.

[36] J. Jacod and P. Protter. *Probability Essentials.* Springer, New York, 2000.

[37] W. C. M. Kallenberg. On moderate and large deviations in multinomial distributions. *The Annals of Statistics*, 13:1554–1580, 1985.

[38] J. H. B. Kemperman. An optimum rate of transmitting information. *The Annals of Mathematical Statistics*, 40:2156–2177, 1969.

[39] S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127, 1967.

[40] S. Kullback, and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[41] F. Liese, and I. Vajda. *Convex Statistical Distances.* Teubner, Leipzig, 1987.

[42] C. McDiarmid. On the method of bounded differences. In *Survey in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

[43] C. Morris. Central limit theorems for multinomial sums. *The Annals of Statistics*, 3:165–188, 1975.

[44] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for the purposes of statistical inference. *Biometrika*, 20A:175–247,264–299, 1928.

[45] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypothese. *Philos. Trans. Roy. Soc. London A*, 231:289–337, 1933.

[46] Ya. Nikitin. *Asymptotic efficiency of nonparametric tests.* Cambridge University Press, Cambridge, 1995.

[47] Pardo, M. C., Pardo, L. and Vajda, I. (2004). Testing homogeneity of independent samples from arbitrary models, *Research Report, No 2104, Institute of Automation and Information Theory of the Czech Academy of Sciences.*

[48] Pardo, L., Pardo, M. C. and Zografos, K. (1999). Homogeneity for multinomial populations based on $\phi$-divergences, *J. Japan Statist. Soc.*, 29, pp. 213–228.

[49] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1901.

[50] M.P. Quine and J. Robinson. Efficiencies of chi-square and likelihood ratio goodness-of-fit tests. *The Annals of Statistics*, 13:727–742, 1985.

[51] C. R. Rao. *Statistical Inference and its Applications.* Wiley, New York, second edition, 1973.

[52] T. Read and N. Cressie. *Goodness-Of-Fit Statistics for Discrete Multivariate Analysis.* Springer-Verlag, New York, 1988.

[53] M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, 3(1):1–14, 1975.

[54] I. N. Sanov. On the probability of large deviations of random variables. *Mat. Sb.*, 42:11-44, 1957. (English translation in *Sel. Transl. Math. Statist. Prob.*, 1:213-244, 1961.)

[55] H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18:434–438, 1947.

[56] R. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley, New York, 1980.

[57] I. S. Shiganov. Refinement of the upper bound of the constant in the central limit theorem (english). *J. Soviet Math.*, 35:2545–2550, 1986.

[58] C. Stein, 1952, published in Chernoff [12].

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

[60] T. J. Sweeting. Speeds of convergence for the multidimensional central limit theorem. *The Annals of Probability*, 5:28–41, 1977.

[61] G. T. Toussaint. Sharper lower bounds for information in term of variation. *IEEE Trans. Information Theory*, IT-21:99-103, 1975.

[62] G. Tusnády. On asymptotically optimal tests. *The Annals of Statistics*, 5:385–393, 1977.

[63] Y. Um and R. Randles. A multivariate nonparametric test among many vectors. *Journal of Nonparametric Statistics*, 13:699–708, 2001.

[64] N. Ushakov. *Selected Topics in Characteristic Functions.* Modern Probability and Statistics. Walter de Gruyter, Berlin, 1999.

[65] I. Vajda. *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, 1989.

[66] I. Vajda. Note on discrimination information and variation. *IEEE Trans. Information Theory*, IT-16:771-773, 1970.

[67] G. S. Watson. On chi–squared goodness–of–fit tests for continuous distributions. *Journal of the Royal Statistical Society, Series B*, 20:44-61, 1958.