mathematics

MDPI

Article

Forward Selection of Relevant Factors by Means of MDR-EFE Method

Alexander Bulinski

Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, Leninskie Gory 1, 119991 Moscow, Russia; alexander.bulinski@math.msu.ru

Abstract: The suboptimal procedure under consideration, based on the MDR-EFE algorithm, provides sequential selection of relevant (in a sense) factors affecting the studied, in general, non-binary random response. The model is not assumed linear, the joint distribution of the factors vector and response is unknown. A set of relevant factors has specified cardinality. It is proved that under certain conditions the mentioned forward selection procedure gives a random set of factors that asymptotically (with probability tending to one as the number of observations grows to infinity) coincides with the "oracle" one. The latter means that the random set, obtained with this algorithm, approximates the features collection that would be identified, if the joint distribution of the features vector and response were known. For this purpose the statistical estimators of the prediction error functional of the studied response are proposed. They involve a new version of regularization. This permits to guarantee not only the central limit theorem for normalized estimators, but also to find the convergence rate of their first two moments to the corresponding moments of the limiting Gaussian variable.

Keywords: feature selection; relevant factors; MDR-EFE method; forward selection; suboptimal procedures; statistical estimators of error functional (of a response); regularized estimators; CLT; convergence of estimators moments

MSC: 62G20; 62H12; 62J02; 62L12

Citation: Bulinski, A. Forward Selection of Relevant Factors by Means of MDR-EFE Method. *Mathematics* **2024**, *12*, 831. https://doi.org/10.3390/math12060831

Academic Editors: Jin-Ting Zhang

Received: 20 January 2024
Revised: 5 March 2024
Accepted: 8 March 2024
Published: 12 March 2024

1. Introduction

This paper is dedicated to the eminent scientist Professor A.S. Holevo, academician of the Russian Academy of Sciences, on occasion of his remarkable birthday.

The classical problem of regression analysis consists in the search for deterministic function f, which, in a certain sense, "well" approximates the observed random variable (response) Y by the value $f(X)$, where $X = (X_1, \dots, X_p)$ is a vector of factors influencing the behavior of Y. This approach was initiated by the works of A.-M. Legendre and K. Gauss. At that time it found application in the processing of astronomical observations. Nowadays one widely uses the methods involving the appropriate choice of unknown real coefficients $\beta_1, \dots, \beta_p$ for a linear model of the form $Y = \sum_{i=1}^{p} \beta_i X_i + \varepsilon$, where ε describes a random error. Clearly, $X_0 = 1$ can be included in the collection of factors, then $Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$. For example, books [1,2] are devoted to regression. The close tasks also arise in observations classification, see, e.g., [3].

Since the end of the 20th century, stochastic models have been studied where the random response Y depended only on some subset of the factors in the set of $X_1, \dots, X_p$. So, in article [4], the LASSO method (Least Absolute Shrinkage and Selection Operator) was introduced, using the idea of regularization (going back to A.N.Tikhonov), which allowed to find factors included with non-zero coefficients in a "sparse" linear model. Somewhat earlier, this approach was used by several authors for the treatment of geophysical data. Generalizations of the mentioned method are considered in monograph [5]. We emphasize that the idea of identifying some of the factors having a principle (in a certain sense) impact

on a response is also intensely developing within the framework of nonlinear models. Such direction of modern mathematical statistics is called Feature Selection (FS), i.e., the choice of features (variables, factors). In this regard, we refer, e.g., to monographs [6–9] and also to reviews [10–14]. In [10] the authors consider filter, wrapper and embedded methods of FS. They concentrate on feature elimination and also demonstrate the application of FS technique on standard datasets. In [11] the modern mainstream dimensionality reduction methods are analyzed including ones for small samples and those based on deep learning. In [12] FS machinery is considered based on filtering methods for detecting the cyber attacks. Survey [13] is devoted to FS methods in machine learning (the structured information is contained in 20 tables). The authors of [14] concentrate on applications of FS to stock market prediction and applications of FS in the analysis of credit risks are considered, e.g., in [15]. Beyond financial mathematics the choice of relevant factors is very important in medicine and biology. For instance, in the field of genetic data analysis there is an extensive research area called GWAS (Genome-Wide Association Studies) aimed at studying the relationships between phenotypes and genotypes, see, e.g., [16,17]. The authors of [18] provide the survey of starting methods used by genetic algorithms. Review [19] is devoted to the FS methods for predicting the risk of diseases. Thus, research in the field of FS is not only of theoretical interest, but also admits various applications.

Note that there are a number of complementary methods for identifying relevant factors. Much attention is paid to those employing the basic concepts of information theory such as entropy, mutual information, conditional mutual information, interaction information, various divergences, etc. Here statistical estimation of information characteristics plays an important role. One can mention, e.g., works [20,21]. In this article, the accent is made on identifying a set of relevant factors in the framework of a certain stochastic model, when the quality of the response approximation is evaluated by means of some metric.

Recall that J.B. Herrick in 1910 described the Sickle cell anemia (HbS). Later it was discovered that all clinical manifestations of the presence of HbS are the consequences of the single change in the B-globin gene. This famous example shows that even the search of a single feature having impact on a disease is reasonable. Nowadays the researchers concentrate on complex diseases provoked by several disorders of the human genome. Even identification of two SNPs (single nucleotide polymorphisms) having impact on a certain disease is of interest, see, e.g., [22].

Now we turn to the description of the studied mathematical model. All the considered random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Let a random variable Y map Ω to some finite set $\mathbb{Y}$. We assume that, for $k \in T := \{1, \ldots, p\}$, a random variable $X_k : \Omega \to M_k$, where M_k is an arbitrary finite set. Then the vector $X = (X_1, \ldots, X_p)$ takes the values in $\mathbb{X} = M_1 \times \ldots \times M_p$. For a set $S = \{i_1, \ldots, i_r\}$, where $1 \leq i_1 < \ldots < i_r \leq p$, we put $X_S := (X_{i_1}, \ldots, X_{i_r})$. Similarly, for $x \in \mathbb{X}$, x_S denotes a vector $(x_{i_1}, \ldots, x_{i_r})$. A collection of indices $S \subset T$ (the symbol $\subset$ is everywhere understood as a non-strict inclusion) is called *relevant* if the following relation holds for any $x \in \mathbb{X}$ and $y \in \mathbb{Y}$:

$$\mathsf{P}(Y = y | X = x) = \mathsf{P}(Y = y | X_S = x_S), \tag{1}$$

whenever $\mathsf{P}(Y = y | X = x) \neq 0$. In this case, the set of factors X_S is called relevant as well. If (1) takes place for some $S = S_0$ then it will be obviously valid for any S containing S_0. Therefore, the natural desire is to identify a set S that satisfies (1) and has cardinality $r < p$ (if such a set other than T exists). Note that there are different definitions of the relevant factors collection, see, e.g., [23,24] and the references therein.

It is assumed that a collection of relevant factors has r elements ($1 \leq r < p$), however, the set S itself, which appears in (1), is unknown and should be identified. We label this assumption as (A). There is no restriction that S satisfying (1) and containing r elements is unique. Usually the joint distribution of (X, Y) is also unknown. Therefore, a statistical estimator of S is constructed based on the first N observations $\xi_N := (\xi^{(1)}, \ldots, \xi^{(N)})$ of a sequence $\xi^{(1)}, \xi^{(2)}, \ldots$, consisting of i.i.d. random vectors, where, for $k \in \mathbb{N}$, $\xi^{(k)} := (X^{(k)}, Y^{(k)})$ has the same distribution as the vector (X, Y).

In 2001, the authors of [25] proposed a method for identifying relevant factors, called MDR (Multifactor Dimensionality Reduction). According to article [26], more than 800 publications were devoted to the development of this method and its applications in the period from 2001 to 2014. Research in this direction has continued over the last decade, see, e.g., [27–29]. In [30], for the binary response Y, a modification of the MDR method was introduced, namely, MDR-EFE (Error Function Estimation), based on statistical estimates of the error functional of the response prediction using the K-fold cross-validation procedure, see also [31]. Later this method was extended in [32] to study the non-binary response.

Recall how the MDR-EFE method is employed. Let a non-random function $f : \mathbb{X} \to \mathbb{Y}$ be used to predict the response Y by the values of the factors vector X. Further we exclude considering the trivial case when $Y = y_0$ with probability one for some $y_0 \in \mathbb{Y}$ (hence, X and Y are independent). The prediction quality is determined by applying the following *error functional*

$$\mathsf{Err}(f) := \mathsf{E}|Y - f(X)|\psi(Y), \tag{2}$$

where a penalty function $\psi : \mathbb{Y} \to \mathbb{R}_+$. The functional Err takes finite values for the discrete X and Y under consideration. The function ψ allows to take into account the importance of approximating a particular value of Y using $f(X)$.

In biomedical research, one often considers the binary response Y characterizing the patient's state of health, say, the value $Y = 1$ corresponds to illness, and $Y = -1$ means that the patient is healthy. In many situations it is more important to consider the disease detection, so the value of 1 is attributed more weight. Of interest is the situation when $\mathbb{Y} = \{-1, 0, 1\}$. Then the value 0 describes some intermediate state of uncertainty ("gray zone"). Following [32], we will consider a more general scheme when the set $\mathbb{Y} := \{-m, \dots, 0, \dots, m\}$ for some $m \in \mathbb{N}$. Lemma 1 in [32] describes for such model all optimal functions f_{opt} that deliver a minimum to the error functional (2). Note that we can suppose that the set of values of Y is strictly contained in $\{-m, \dots, m\}$, i.e., some values are accepted with zero probability. For such y, we assume that $\psi(y) = 0$. Thus, it is possible to study Y taking values in an arbitrary finite subset of $\mathbb{Z}$. In order to simplify the notation, we further consider $\mathsf{P}(Y = y) > 0$ for all $y \in \mathbb{Y} = \{-m, \dots, m\}$.

It is proved that in the framework of model (1) the relation $f_{opt} = f^S$ is valid, where, for $x \in \mathbb{X}$ and $U \subset T$, $f^U(x) = f(x_U)$ and a function f is constructed in a due way. At the same time, for any $U \subset T$ such that $\sharp U = \sharp S$ ($\sharp$ denotes the cardinality of a finite set) and S appearing in (1), the following inequality is true:

$$\mathsf{Err}(f^S) \le \mathsf{Err}(f^U). \tag{3}$$

For $U \subset T$, the function f^U is introduced further. It depends on the joint distribution of (X, Y) which is usually unknown. Thus we use observations $\xi_N = \{(X^{(j)}, Y^{(j)}), j = 1, \dots, N\}$ for statistical estimates of the functional $\mathsf{Err}(f^U)$, where $U \subset T$, and then select as an estimator of S the set U on which the minimum of the corresponding statistical estimate is attained. This approach is described in the next section of the article.

We underline that consideration of all subsets (of the set T) having the cardinality r in the mentioned comparison procedure (involving regularized estimators, as explained in Section 2) for statistical estimates of the error functional is practically unfeasible, when p is large and r is moderately large. Therefore, a number of suboptimal methods of sequential feature selection have emerged. Such methods are used in various approaches to identify sets of relevant factors.

Mainly, one aims either to sequentially add indexes at each step of the algorithm for constructing a statistical estimator of a set S appearing in (1), or to sequentially exclude features from the general set T. In [33], algorithms of forward selection, i.e., sequential addition of indexes to the initial set, based on information theory, are considered. The authors of [33] show that the various algorithms employed can be interpreted as procedures based on proper approximations of the certain objective function. In [34] the principle attention is paid to simple models describing the phenomenon of epistasis observed in

genetics, when individual factors do not affect the response, and some combinations of them lead to essential effects (in statistics one says "synergy interaction" of factors). Besides we also demonstrated that a number of well-known algorithms, for instance, mRMR (Minimum Redundancy Maximum Relevance) using mutual information and/or interaction information with a sequential procedure for selecting relevant factors can lead to the identification of the desired set with probability which is negligibly small. In [35] a variant is proposed for sequential (forward) application of the MDR-EFE method within the binary response model involving the naive Bayesian classifier scheme. The latter means that, for any $y \in \{-1, 1\}$ and all $x \in \mathbb{X}$, the following relation holds:

$$\mathsf{P}(X = x|Y = y) = \prod_{k=1}^{p} \mathsf{P}(X_k = x_k|Y = y). \tag{4}$$

In other words, the factors $X_1, \ldots, X_p$ are conditionally independent for a given response Y. In [35] the joint distribution of X and Y was assumed known.

The principle goal of our work is to derive, for a non-binary, in general, random response, the probability that a sequential selection of features based on the (forward) application of the MDR-EFE method, without assuming the validity of (4), leads to identifying a suboptimal set that would be constructed by means of the same method from observations with a known joint distribution of the response and the vector of factors.

This result builds on the central limit theorem (CLT) for statistical estimates of the prediction error functional for a possibly non-binary response, proved in [32], which extends the CLT for the binary response model studied by the author previously. In addition, for the purposes of this work, we found the convergence rate of the first two moments of the considered statistics to the corresponding moments of the limiting Gaussian variable as the number of observations tends to infinity.

The article has the following structure. Section 2 describes statistical estimates of the error functional (for a response prediction) based on the MDR-EFE method. We also introduce the regularized versions of these estimators. In Section 3, the convergence rate of the first two moments of the regularized estimators of the error functional to the corresponding moments of the limiting Gaussian variable is established. Section 4 contains the main result related to the forward selection of relevant factors. The concluding remarks are given in Section 5. The proof of elementary Lemma 2 is provided in Appendix A for completeness of exposition.

2. Error Functional Estimators

Consider, in general, a non-binary response, i.e., let $\mathbb{Y} := \{-m, \ldots, 0, \ldots, m\}$ for some $m \in \mathbb{N}$. In the framework of the introduced discrete model, Lemma 1 of [32] gives a complete description of the class of optimal functions f_{opt} providing the minimum error $\mathsf{Err}(f)$, determined by (2), in the class of all functions $f : \mathbb{X} \to \mathbb{Y}$. To define such a function (included in the optimal class) for $x \in \mathbb{X}$, we deal with a vector $w(x)$ having components

$$w_y(x) := \psi(y)\mathsf{P}(Y = y, X = x), \;\; y \in \mathbb{Y}.$$

It can be easily seen that

$$\mathsf{Err}(f) = \sum_{y,z \in \mathbb{Y}} |y - z|\psi(y)\mathsf{P}(Y = y, f(X) = z) = \sum_{z \in \mathbb{Y}} \sum_{x \in A_z} w^{\top}(x)q(z), \tag{5}$$

where $A_z := \{x \in \mathbb{X} : f(x) = z\}$, $q(z)$ is a column of $(2m + 1) \times (2m + 1)$ matrix Q having elements $q_{y,z} := |y - z|$ (the element $q_{-m,-m}$ is located in the upper left corner of the matrix Q), $\top$ stands for the transposition of column vectors. In other words, one employs in (5) the scalar product of the vectors $w(x)$ and $q(z)$. Thus, search for an optimal function f_{opt} means finding the partition of $\mathbb{X}$ into such sets A_z, $z \in \mathbb{Y}$, that provide the minimum value

of the right-hand side of (5). Note also that, according to Formula (13) of [32], the error of response prediction can be written as follows:

$$\mathsf{Err}(f) = \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \psi(y)\mathsf{P}(Y = y, |f(X) - y| > i). \tag{6}$$

Let, for $y \in \mathbb{Y}$, the vector $\Delta(y)$ have the first $m + y$ components equal to 1, and the remaining $m - y + 1$ components equal to (-1). For any $x \in \mathbb{X}$, we introduce a vector $L(x)$ with $2m$ components having the form

$$L_y(x) := w^\top(x)\Delta(y), \;\; y \in \mathbb{Y}, \; y > -m. \tag{7}$$

According to formula (11) of [32] one infers that

$$f_{opt}(x) = y \Longleftrightarrow \begin{cases} L_{-m+1}(x) \ge 0, & y = -m, \\ L_{y+1}(x) \ge 0, \; L_y(x) < 0, & y \ne \pm m, \\ L_m(x) < 0, & y = m. \end{cases} \tag{8}$$

The joint distribution of (X, Y) is, in general, unknown. Therefore, the optimal function f_{opt} cannot be found in practice, so an algorithm is used to predict it, i.e., to approximate by means of specified statistical estimators. The response prediction algorithm is defined as a function $\widehat{f}_{PA} = \widehat{f}_{PA}(x, \xi(W))$ given for $x \in \mathbb{X}$ and a set of observations

$$\xi(W) := \{\xi^{(j)} = (X^{(j)}, Y^{(j)}), j \in W\}, \;\; W \subset \mathbb{N}, \;\; \sharp W < \infty. \tag{9}$$

The function $\widehat{f}_{PA}$ takes values in the set $\mathbb{Y}$. It is assumed that the value of $\widehat{f}_{PA}(x, \xi(W))$ becomes close, in a certain sense, to $f(x)$ for x in a specified subset of the set $\mathbb{X}$ when W is sufficiently "massive". More precisely, we consider a family of functions $\widehat{f}_{PA}$ that depend on sets $\xi(W)$ of different cardinalities, but we will not complicate the notation. Consider $M = \{x \in \mathbb{X} : \mathsf{P}(X = x) > 0\}$. For $x \in \mathbb{X}$, $U \subset T$ and $y \in \mathbb{Y}$, introduce a vector $w^U(x)$ with components

$$w_y^U(x) := \begin{cases} \psi(y)\mathsf{P}(Y = y, X_U = x_U), & x \in M, \\ 0, & x \notin M. \end{cases}$$

Set

$$L_y^U(x) := (w^U(x))^\top \Delta(y), \;\; y \in \mathbb{Y}, \;\; y > -m. \tag{10}$$

For $U \subset T$, let f^U be defined by means of a counterpart of formula (8), where $L_y^U(x)$ is now written instead of $L_y(x)$. Then, according to Section 5 of [32] (the notation α is used there instead of U), in the framework of model (1), the optimal function $f_{opt} = f^S$, where S appears in (1) and $\sharp S = r$. Therefore relation (3) is valid for f^U corresponding to any $U \subset T$ with $\sharp U = r$ (the assumption (A) holds).

To introduce an algorithm for predicting the function f^U, we employ statistical estimators of the penalty function ψ, as well as the values $L_y^U(x)$, where $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $y > -m$. Consider

$$\psi(y) := 1/\mathsf{P}(Y = y), \;\; \text{where} \;\; \mathsf{P}(Y = y) > 0, \;\; y \in \mathbb{Y}. \tag{11}$$

In the case of a binary response, such a choice of the penalty function was proposed in [36], the justification for this choice is given in [31], see also Section 4 in [32]. For the specified function $\psi(y)$ and observations $\xi(W)$, where the finite set $W \subset \mathbb{N}$, we use

$$\widehat{\psi}(y, \xi(W)) := \begin{cases} \frac{1}{\widehat{P}(y, \xi(W))}, & \widehat{P}(y, \xi(W)) \ne 0, \\ 0, & \widehat{P}(y, \xi(W)) = 0, \end{cases} \tag{12}$$

where the frequency estimator of a probability $\mathsf{P}(Y = y)$ has the form

$$\widehat{P}(y, \xi(W)) := \frac{1}{\sharp W} \sum_{j \in W} \mathbb{I}\{Y^{(j)} = y\}, \quad y \in \mathbb{N}. \tag{13}$$

It is not difficult to see that the strong law of large numbers for arrays of random variables (see, e.g., [37]) entails for finite sets $W_N \subset \mathbb{N}$, such that $\sharp W_N \to \infty$, the relation

$$\widehat{\psi}(y, \xi(W_N)) \to \psi(y) \ \text{ a.s.}, \quad N \to \infty. \tag{14}$$

Let the prediction algorithm $\widehat{f}^{\,U}_{PA}(x, \xi(W_N))$ of a function $f^U(x)$ be constructed by means of formula (8) analogue, where, for $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $y > -m$, and $W_N \subset \{1, \dots, N\}$, one uses now statistical estimators $\widehat{L}^{U,W_N}_y(x)$ of functions $L^U_y(x)$ introduced in (10). Namely, let us define the following random variables:

$$\widehat{w}^{U,W_N}_y(x) := \widehat{\psi}(y, \xi(W_N)) \frac{1}{\sharp W_N} \sum_{j \in W_N} \mathbb{I}\{Y^{(j)} = y, X^{(j)}_U = x_U\}, \quad y \in \mathbb{Y},$$

where $\widehat{\psi}(y, \xi(W_N))$ is an estimator of $\psi(y)$ appearing in (12). For $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $y > -m$, set

$$\widehat{L}^{U,W_N}_y(x) := \widehat{w}^{U,W_N}_y(x)^\top \Delta(y).$$

Replace the value $L_y(x)$ in (8) by $\widehat{L}^{U,W_N}_y(x)$. Then one can claim that

$$\widehat{f}^{\,U}_{PA}(x, \xi(W_N)) = y \Longleftrightarrow \begin{cases} \widehat{L}^{U,W_N}_y(x) \geq 0, & y = -m, \\ \widehat{L}^{U,W_N}_{y+1}(x) \geq 0, \ \widehat{L}^{U,W_N}_y(x) < 0, & y \neq \pm m, \\ \widehat{L}^{U,W_N}_y(x) < 0, & y = m. \end{cases} \tag{15}$$

For $K \in \mathbb{N}$, $K > 1$, we take a partition of a set $\{1, \dots, N\}$ into subsets

$$D_k(N) := \{(k-1)[N/K] + 1, \dots, k[N/K]\mathbb{I}\{k < K\} + N\mathbb{I}\{K = N\}\}, \tag{16}$$

here $k = 1, \dots, K$, $[a]$ is an integer part of a number $a \in \mathbb{R}$, $\mathbb{I}\{A\}$ is an indicator of a set A. These sets are applied in the K-fold cross-validation procedure increasing the stability of statistical inference (cross-validation procedure is studied, e.g., in [38]). Following [32], the estimator of the functional $\mathsf{Err}(f^U)$, i.e., a statistical estimator of the prediction error functional for a function f^U and observations $\xi_N := \xi(\{1, \dots, N\})$, involving the K-fold cross-validation procedure, is given by the formula:

$$\widehat{Err}_{K,N}(f^U) := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\leq m} \frac{1}{K} \sum_{k=1}^{K} \widehat{\psi}(y, \xi(D_k(N)))$$

$$\times \frac{1}{\sharp D_k(N)} \sum_{j \in D_k(N)} \mathbb{I}\{Y^{(j)} = y, |\widehat{f}^{\,U}_{PA}(X^{(j)}, \xi(\overline{D}_k(N))) - y| > i\}, \tag{17}$$

where $\overline{D}_k(N) := \{1, \dots, N\} \setminus D_k(N)$ and $\widehat{\psi}(y, \xi(D_k(N)))$ are evaluated according to (12) for $W_N = D_k(N)$, $k = 1, \dots, K$. The estimator (17) is a natural statistical analogue of the error functional (2) written in the form (6) when one employs the K-cross-validation procedure. Namely, instead of $\psi(y)$ we apply its statistical estimator of the type (12) and instead of f we use its approximation by means of prediction algorithm based on the part $D_k(N)$ of observations. To obtain the statistical estimators of the probability appearing in Formula (6) we write the corresponding average of indicator functions. One employs also the averaging over different parts of observations.

By Theorem 2 of [32], if $S = \{i_1, \ldots, i_r\}$ is a set of relevant factors, i.e., (1) holds, then, for each $\varepsilon > 0$ and any set $U = \{m_1, \ldots, m_r\} \subset T$, the following inequality takes place almost sure for all N large enough:

$$\widehat{Err}_{K,N}(f^S) \leq \widehat{Err}_{K,N}(f^U) + \varepsilon. \tag{18}$$

Thus, it is natural to consider all subsets $U = \{m_1, \ldots, m_r\} \subset T$ and choose as a statistical estimator of a relevant collection of indices $(i_1, \ldots, i_r)$ a set U on which the minimum of $\widehat{Err}_{K,N}(f^U)$ is attained. Here we also note that, for the study of asymptotic properties of the error functional, the regularization of the prediction algorithm by means of a sequence of positive numbers $(\varepsilon_N)_{N\in\mathbb{N}}$ such that $\varepsilon_N \to 0$, as $N \to \infty$, plays an important role. Namely, for $W_N \subset \{1, \ldots, N\}$, we define

$$\widehat{f}^{\,U}_{PA,\varepsilon_N}(x, \xi(W_N)) = y \Longleftrightarrow \begin{cases} \widehat{L}_y^{U,W_N}(x) + \varepsilon_N \geq 0, & y = -m, \\ \widehat{L}_{y+1}^{U,W_N}(x) + \varepsilon_N \geq 0,\ \widehat{L}_y^{U,W_N}(x) + \varepsilon_N < 0, & y \neq \pm m, \\ \widehat{L}_y^{U,W_N}(x) + \varepsilon_N < 0, & y = m. \end{cases} \tag{19}$$

As in article [32], we assume that

$$\varepsilon_N \to 0+, \quad \sqrt{N}\varepsilon_N \to \infty, \quad N \to \infty. \tag{20}$$

Now we introduce a statistical estimator $\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)$ using an analogue of Formula (17), where one employs $\widehat{f}^{\,U}_{PA,\varepsilon_N}$ instead of $\widehat{f}^{\,U}_{PA}$. For the regularized statistical estimators, as mentioned in [32], the analogue of Formula (18) holds. In [32], for estimators $\widehat{f}^{\,U}_{PA,\varepsilon_N}$ constructed when condition (20) is met, the CLT is established. In the next section we apply a slightly different regularization for the error functional estimates, which will permit us to specify the convergence rate of the first two moments of these estimators to corresponding moments of the limiting Gaussian variable. This result is not only of independent interest, but is also applied in Section 4.

3. Asymptotic Behavior of the First Two Moments of Statistical Estimators of the Error Functional

As noted in Section 2, we will use the penalty function (11). Therefore, for $W_N = D_k(N)$, as a strongly consistent estimator $\widehat{\psi}(y, D_k(N))$ of $\psi(y)$ we will employ the variable appearing in (12), denoted below as $\widehat{\psi}_{N,k}(y)$, where $y \in \mathbb{Y}$, $k = 1, \ldots, K$, $N \in \mathbb{N}$. Recall that the estimator $\widehat{\mathsf{Err}}_{K,N}(f^U)$ is defined by formula (2). If the regularized version $\widehat{f}_{PA,\varepsilon_N}$ is substituted into this estimator instead of $\widehat{f}^{\,U}_{PA}$, where $x \in \mathbb{X}$ and $N \in \mathbb{N}$, then the notation $\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)$ is used. We will apply the following Corollary 3 of [32] established in the framework of a model satisfying (1).

Theorem 1 ([32]). *Let U be an arbitrary subset of T having the cardinality r, the function f^U be defined after formula (10), $\widehat{f}^{\,U}_{PA,\varepsilon_N}$ appear in (19) for observations ξ_N, and the sequence $(\varepsilon_N)_{N\in\mathbb{N}}$ satisfy condition (20). Then*

$$\sqrt{N}\left(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U) - \mathsf{Err}(f^U)\right) \xrightarrow{\mathcal{D}} Z \sim N(0, \sigma^2(U)), \quad N \to \infty, \tag{21}$$

and in this case $\sigma^2(U)$ is the variance of a random variable

$$V(U) := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\leq m} \frac{\mathbb{I}\{Y = y\}}{\mathsf{P}(Y = y)} (\mathbb{I}\{|f^U(X) - y| > i\} - \mathsf{P}(|f^U(X) - y| > i | Y = y)). \tag{22}$$

It is known that the convergence in distribution of random variables, in general, does not ensure the convergence of their moments even when the moments exist. We will manage to establish the convergence rate of the first two moments of the error functional

statistical estimators to the corresponding moments of the limit random variable. For this purpose we slightly strength the condition of estimates regularization. We require that a sequence $(\varepsilon_N)_{N\in\mathbb{N}}$ satisfies the following condition:

$$\varepsilon_N \to 0+, \quad \frac{\varepsilon_N\sqrt{N}}{\sqrt{\log\frac{1}{\varepsilon_N}}} \to \infty, \quad N\to\infty. \tag{23}$$

Clearly, (23) implies the validity of (20). Relation (23) holds if one takes $\varepsilon_N = N^{-\delta}$, $N\in\mathbb{N}$, where $\delta\in(0,1/2)$.

Lemma 1. *Let condition (23) be met. Then, for every $K\in\mathbb{N}$, $K>1$, and any $U\subset T$, the statistical estimators $\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)$ satisfy the following relation:*

$$N\,\mathsf{E}(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)-\mathsf{Err}(f^U))^2 \to \sigma^2(U), \quad N\to\infty, \tag{24}$$

where $\sigma^2(U)=\mathsf{var}\,V(U)$ and $V(U)$ is introduced in formula (22).

Proof of Lemma 1. Let us fix an arbitrary set $U\subset T$. For each $N\in\mathbb{N}$ one has

$$\begin{aligned}\mathbb{Z}_N \;:=\; & \sqrt{N}\Big(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)-\mathsf{Err}(f^U)\Big) = \sqrt{N}(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)-\widehat{T}_N(f^U)) \\ +\; & \sqrt{N}(\widehat{T}_N(f^U)-T_N(f^U))+\sqrt{N}(T_N(f^U)-\mathsf{Err}(f^U)),\end{aligned} \tag{25}$$

where

$$T_N(f^U) := \sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{1}{K}\sum_{k=1}^{K}\frac{\psi(y)}{\sharp D_k(N)}\sum_{j\in D_k(N)}\mathbb{I}\{Y^{(j)}=y, |f^U(X^{(j)})-y|>i\}, \tag{26}$$

$$\widehat{T}_N(f^U) := \sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{1}{K}\sum_{k=1}^{K}\frac{\widehat{\psi}_{N,k}(y)}{\sharp D_k(N)}\sum_{j\in D_k(N)}\mathbb{I}\{Y^{(j)}=y, |f^U(X^{(j)})-y|>i\}, \tag{27}$$

$\widehat{\psi}_{N,k}(y)$ are defined by means of (12) for $W_N=D_k(N)$, $k=1,\dots,K$, $N\in\mathbb{N}$. The proof is divided into several steps.

Step 1 . At first we consider

$$R_N := \sqrt{N}(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)-\widehat{T}_N(f^U)), \quad N\in\mathbb{N}.$$

To simplify the notation, we do not write that R_N also depends on K, ξ_N and ε_N. Our aim is to show that if (23) holds then

$$\mathsf{E}R_N^2\to 0 \;\text{ as }\; N\to\infty. \tag{28}$$

In the light of formula (71) of [32], under condition (20) the following relation is valid:

$$R_N \xrightarrow{\mathsf{P}} 0, \quad N\to\infty. \tag{29}$$

Taking into account (29), by Theorem 5.4 of [39], relation (28) holds if (and only if) the sequence $(R_N^2)_{N\in\mathbb{N}}$ is uniformly integrable. Due to theorem by De La Vallé - Poussin (see, e.g., Theorem 1.3.4 of [40]) it is sufficient to verify that

$$\sup_{N\in\mathbb{N}}\mathsf{E}(R_N^4)<\infty.$$

For $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $i \in \mathbb{Z}_+$, $k = 1, \ldots, K$ and $N \in \mathbb{N}$ we introduce the following random variables:

$$F_{N,k}^{(i)}(x,y) = \mathbb{I}\{|\widehat{f}_{PA,\varepsilon_N}^{U}(x, \xi(\overline{D}_{N,k})) - y| > i\} - \mathbb{I}\{|f^{U}(x) - y| > i\}, \tag{30}$$

$$\mathbb{S}_k(i,y) := \frac{1}{\sharp D_k(N)} \sum_{j \in D_k(N)} \mathbb{I}\{Y^{(j)} = y\} F_{N,k}^{(i)}(X^{(j)}, y), \tag{31}$$

where, for $W \subset \mathbb{N}$, $\xi(W)$ is defined by Formula (9). Write $R_N = U_{N,1} + U_{N,2}$, here

$$\begin{aligned} U_{N,1} &:= \sqrt{N}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\psi(y)\mathbb{S}_k(i,y)\right), \\ U_{N,2} &:= \sqrt{N}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}(\widehat{\psi}_{N,k}(y) - \psi(y))\mathbb{S}_k(i,y)\right). \end{aligned}$$

Now note that, for any real numbers $a_1, \ldots, a_v$, every $v \in \mathbb{N}$ and an arbitrary $\gamma > 1$, the Hölder inequality implies that

$$\left(\sum_{r=1}^{v}|a_r|\right)^{\gamma} \le v^{\gamma-1}\sum_{r=1}^{v}|a_r|^{\gamma}. \tag{32}$$

Evidently, (32) is true for $\gamma = 1$ as well. Consequently, we get

$$R_N^4 \le 8(U_{N,1}^4 + U_{N,2}^4), \quad N \in \mathbb{N}. \tag{33}$$

Clearly, for all $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $W_N \subset \{1, \ldots, N\}$ and $N \in \mathbb{N}$, one has

$$\widehat{L}_{y,\varepsilon_N}^{U,W_N}(x) := \widehat{L}_y^{U,W_N}(x) + \varepsilon_N = L_y^{U}(x) + (\widehat{w}_y^{U,W_N}(x) - w_y(x))^{\top}\Delta(y) + \varepsilon_N, \tag{34}$$

where the functions appearing in (34) were introduced in Section 2. For any $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, the inequalities $L_y^U(x) \ge 0$, $L_{y+1}^U(x) < 0$ are satisfied if and only if, for arbitrary $\delta_N(x,y;U) > 0$ such that $\delta_N(x,y;U) \to 0$, as $N \to \infty$, and all sufficiently large $N \in \mathbb{N}$, the following inequalities are valid: $L_y^U(x) + \delta_N(x,y;U) > 0$, $L_{y+1}^U(x) + \delta_N(x,y;U) < 0$ (the analogous statement is true for inequalities corresponding to coordinates $y = m$ and $y = -m$ in Formula (19)). Obviously,

$$\begin{aligned} &|(\widehat{w}_y^{U,W_N}(x) - w_y(x))^{\top}\Delta(y)| \\ &\le |\widehat{\psi}(y, \xi(W_N)) - \psi(y)| + \psi(y)\left|\frac{1}{\sharp W_N}\sum_{q \in W_N}\mathbb{I}\{X_U^{(q)} = x_U, Y^{(q)} = y\} - \mathsf{P}(X_U = x_U, Y = y)\right|, \end{aligned}$$

where $\widehat{\psi}(y, \xi(W_N))$ is defined in (12). One has

$$\begin{aligned} &\sum_{x_U}\left(\frac{1}{\sharp W_N}\sum_{q \in W_N}\mathbb{I}\{X_U^{(q)} = x_U, Y^{(q)} = y\} - \mathsf{P}(X_U = x_U, Y = y)\right) \\ &= \frac{1}{\sharp W_N}\sum_{q \in W_N}\mathbb{I}\{Y^{(q)} = y\} - \mathsf{P}(Y = y) \\ &= \widehat{P}(y, \xi(W_N)) - \mathsf{P}(Y = y). \end{aligned} \tag{35}$$

For $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $W_N \subset \{1, \ldots, N\}$ and $N \in \mathbb{N}$, consider the following event

$$A_{W_N}(x,y) = \left\{ \left| \frac{1}{\sharp W_N} \sum_{q \in W_N} \mathbb{I}\{X_U^{(q)} = x_U, Y^{(q)} = y\} - \mathsf{P}(X_U = x_U, Y = y) \right| \leq \frac{p_0^2 \varepsilon_N}{8 \sharp \mathbb{X}} \right\}, \quad (36)$$

where $p_0 = \min_{y \in \mathbb{Y}} \mathsf{P}(Y = y)$ (we assumed that $\mathsf{P}(Y = y) > 0$ for $y \in \mathbb{Y}$). More precisely one can write $A_{W_N}(x,y) = A_{W_N}(x,y,U;\{(X^{(q)}, Y^{(q)}), q \in W_N\})$. We will not include a set U in the list of arguments since this set is fixed. Then, for $\omega \in A_{W_N}(x,y)$, in view of (35), we get

$$\left| \widehat{P}(y, \xi(W_N)) - \mathsf{P}(Y = y) \right| \leq \frac{p_0^2 \varepsilon_N}{8}. \quad (37)$$

Then by virtue of (37), for any $y \in \mathbb{Y}$ and all N large enough, i.e., for $N \geq N_0(\mathbb{Y}, (\varepsilon_N)_{N \in \mathbb{N}})$, one has

$$\widehat{P}(y, \xi(W_N)) \geq \mathsf{P}(Y = y) - \frac{p_0^2 \varepsilon_N}{8} \geq \mathsf{P}(Y = y) - \frac{\varepsilon_N}{8} > \frac{\mathsf{P}(Y = y)}{2} > 0,$$

and hence the following relation holds

$$|\widehat{\psi}(y, \xi(W_N)) - \psi(y)| = \frac{|\widehat{P}(y, \xi(W_N)) - \mathsf{P}(Y = y)|}{\widehat{P}(y, \xi(W_N)) \mathsf{P}(Y = y)} \leq \frac{\frac{p_0^2 \varepsilon_N}{8}}{\frac{\mathsf{P}(Y=y)^2}{2}} \leq \frac{\varepsilon_N}{4}. \quad (38)$$

Thus if $\omega \in A_{W_N}(x,y)$, where $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, then according to (36) and (38), for all N large enough, we can write

$$|(\widehat{w}_y^{U,W_N}(x) - w_y(x))^\top \Delta(y)| \leq \frac{\varepsilon_N}{4} + \left(\frac{1}{p_0} \right) \frac{p_0^2 \varepsilon_N}{8 \sharp \mathbb{X}} \leq \frac{\varepsilon_N}{2}.$$

Taking into account that the sets $\mathbb{X}$ and $\mathbb{Y}$ have finite cardinalities, we ascertain that, for any $x \in \mathbb{X}$, $y \in \mathbb{Y}$ and all N large enough, for $\omega \in A_{W_N}(x,y)$, one has

$$\widehat{f}_{PA,\varepsilon_N}^{U,W_N}(x) = f^U(x). \quad (39)$$

Consequently, for any $x \in \mathbb{X}$, $y \in \mathbb{Y}$, $i = 0, 1, \ldots, 2m-1$, $\omega \in A_{W_N}(x,y)$, where $W_N = \overline{D}_k(N)$, $k = 1, \ldots, K$, for all N large enough (i.e., $N \geq N_1$), the following inequality holds:

$$F_{N,k}^{(i)}(x,y) \mathbb{I}\{A_{\overline{D}_k(N)}(x,y)\} = 0. \quad (40)$$

Applying (32) we come to the inequality

$$|U_{N,1}|^4 \leq N^2 \frac{(2m)^6}{K} \sum_{k=1}^{K} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\leq m} \psi(y)^4 \left(\frac{1}{\sharp D_k(N)} \sum_{j \in D_k(N)} \mathbb{I}\{Y^{(j)} = y\} F_{N,k}^{(i)}(X^{(j)}, y) \right)^4.$$

Let $\widetilde{\Sigma}$ denote the summation over all $x_j \in \mathbb{X}$ for $j \in D_k(N)$. For $N \geq N_1$ one has

$$\begin{aligned} & \mathsf{E} \left(\sum_{j \in D_k(N)} \mathbb{I}\{Y^{(j)} = y\} F_{N,k}^{(i)}(X^{(j)}, y) \right)^4 \\ = \; & \mathsf{E} \left(\widetilde{\Sigma} \left(\sum_{j \in D_k(N)} \mathbb{I}\{Y^{(j)} = y\} F_{N,k}^{(i)}(x_j, y) \right)^4 \mathbb{I}\left\{ \bigcap_{j \in D_k(N)} \{X^{(j)} = x_j\} \right\} \right) \end{aligned}$$

$$= \mathsf{E}\left(\widetilde{\Sigma}\left(\sum_{j\in D_k(N)}\mathbb{I}\{Y^{(j)}=y\}F^{(i)}_{N,k}(x_j,y)\mathbb{I}\{\overline{A}_{\overline{D}_k(N)}(x_j,y)\}\right)^4\mathbb{I}\left\{\bigcap_{j\in D_k(N)}\{X^{(j)}=x_j\}\right\}\right)$$

$$= \mathsf{E}\left(\sum_{j\in D_k(N)}\mathbb{I}\{Y^{(j)}=y\}F^{(i)}_{N,k}(X^{(j)},y)\mathbb{I}\{\overline{A}_{\overline{D}_k(N)}(X^{(j)},y)\}\right)^4$$

$$\le \mathsf{E}\left(\sum_{j\in D_k(N)}\mathbb{I}\{\overline{A}_{\overline{D}_k(N)}(X^{(j)},y)\}\right)^4,$$

here we employ (40) and take into account that $|F^{(i)}_{N,k}(x,y)|\le 1$. We see that

$$|U_{N,1}|^4 \le N^2\frac{(2m)^6}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{\psi(y)^4}{(\sharp D_k(N))^4}\Big(\sum_{j\in D_k(N)}\mathbb{I}\{\overline{A}_{\overline{D}_N(k)}(X^{(j)},y)\}\Big)^4. \quad (41)$$

For $W_N\subset\{1,\dots,N\}$, $y\in\mathbb{Y}$ and $j=1,\dots,N$, introduce the functions

$$g_{W_N}(X^{(j)},y)=\mathbb{I}\{\overline{A}_{W_N}(X^{(j)},y)\}=\mathbb{I}\{\overline{A}_{W_N}(X^{(j)},y;\{(X^{(q)},Y^{(q)}),q\in W_N\})\}.$$

It is known (see, e.g., formula (15) in Chap. VI of [41]) that if a bounded Borel function $g:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$, ξ and ζ are independent random vectors taking values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then

$$\mathsf{E}(g(\xi,\zeta)|\zeta=z)=\mathsf{E}g(\xi,z),\ \ z\in\mathbb{R}^n.$$

Due to independence of $(X^{(j)},Y^{(j)})$, $j\in\mathbb{N}$, we can apply the lemma on grouping random vectors (see, e.g., [42], p. 28) to get the relation

$$\mathsf{E}\Big(\Big(\sum_{j\in D_k(N)}g_{\overline{D}_k(N)}(X^{(j)},y;(X^{(q)},Y^{(q)}),q\in\overline{D}_k(N))\Big)^4\Big|(X^{(q)},Y^{(q)})=(x_q,y_q),q\in\overline{D}_k(N)\Big)$$

$$=\mathsf{E}\Big(\sum_{j\in D_k(N)}g_{\overline{D}_k(N)}(X^{(j)},y;(x_q,y_q)),q\in\overline{D}_N(k))\Big)\Big)^4.$$

By the Rosenthal inequality (see, e.g., Theorem 2.9 of [43]), for independent centered random variables $Z_1,\dots,Z_v$, having $\mathsf{E}|Z_j|^t<\infty$ for some $t\in[2,\infty)$ and each $j=1,\dots,v$, one has

$$\mathsf{E}\Big|\sum_{j=1}^{v}Z_j\Big|^t\le C(t)\Big(\sum_{j=1}^{v}\mathsf{E}|Z_j|^t+\Big(\sum_{j=1}^{v}\mathsf{E}Z_j^2\Big)^{\frac{t}{2}}\Big), \quad (42)$$

where $C(t)>0$ depends on t but does not depend on v and distributions of variables Z_j, $j=1,\dots,v$.

Set $\eta^{(j)}_{N,k}:=g_{\overline{D}_k(N)}(X^{(j)},y;\{(x_q,y_q)),q\in\overline{D}_N(k)\})$, $j\in\mathbb{N}$. Note that $0\le\eta^{(j)}_{N,k}\le 1$ for all $j\in D_N(k)$. Then according to (42) we come to the inequality

$$\mathsf{E}\left(\sum_{j\in D_k(N)}(\eta^{(j)}_{N,k}-\mathsf{E}\eta^{(j)}_{N,k})\right)^4\le C(\sharp D_k(N))^2,$$

where $k=1,\dots,K$ and $C=2C(4)$. Hence, applying (32) for $\gamma=4$ and $v=2$, one has

$$\mathsf{E}\left(\sum_{j\in D_k(N)}\eta^{(j)}_{N,k}\right)^4\le 8\left(C(\sharp D_k(N))^2+\left(\sum_{j\in D_k(N)}\mathsf{E}\eta^{(j)}_{N,k}\right)^4\right)$$

$$\leq 8C(\sharp D_k(N))^2 + 8(\sharp D_k(N))^4 \max_{j\in D_k(N)} (\mathsf{E}\eta_{N,k}^{(j)})^4.$$

Evidently, we can write

$$\mathsf{E}(\eta_{N,k}^{(j)}) = \mathsf{P}(\overline{A}_{\overline{D}_k(N)}(X^{(j)}, y; \{(x_q, y_q)), q \in \overline{D}_k(N)\}).$$

Let $M_k = \sharp\overline{D}_k(N)$, where $M_k = M_k(N)$, $k = 1, \ldots, K$. Set $\zeta_q = \mathbb{I}\{X_U^{(q)} = x_U, Y^{(q)} = y\}$, where $q \in \overline{D}_k(N)$, $\sigma_0^2 = \mathsf{var}\,\zeta_q$. Clearly, ζ_q depends on x_U, y and U. Random variables ζ_q are identically distributed for $q \in \mathbb{N}$. Therefore $\sigma_0^2 = \sigma_0^2(U, x, y)$, but does not depend on q. If $\sigma_0^2 = 0$, then the variables ζ_q are a.s. equal to some constant. According to (36), an event $\overline{A}_{\overline{D}_k(N)}(X^{(j)}, y; \{(x_q, y_q)), q \in \overline{D}_k(N)\})$ occurrence means that the variable which is equal to zero a.s. turns greater than $(p_0^2\varepsilon_N)/(8\sharp\mathbb{X})$. Therefore, in the degenerate case one has

$$\mathsf{P}(\overline{A}_{\overline{D}_k(N)}(X^{(j)}, y; (x_q, y_q)), q \in \overline{D}_k(N))) = 0$$

and $\mathsf{E}\eta_{N,k}^{(j)} = 0$ for all $j = 1, \ldots, N$. Consider now the case when $\sigma_0^2 > 0$. Then we get

$$\mathsf{P}(\overline{A}_{\overline{D}_k(N)}(X^{(j)}, y; \{(x_q, y_q), q \in \overline{D}_k(N)\}) = \mathsf{P}\left(\frac{\sum_{q\in\overline{D}_k(N)}(\zeta_q - \mathsf{E}\zeta_q)}{\sigma_0\sqrt{M_k}} > \frac{p_0^2\sqrt{M_k}\varepsilon_N}{8\sharp\mathbb{X}\sigma_0}\right),$$

where p_0 appeared in (36).

Now we employ the Berry-Esseen estimate of the convergence rate in CLT for i.i.d. random variables. Let $Z_1, \ldots, Z_v$ be i.i.d. random variables such that $\mathsf{E}Z_1 = 0$, $\mathsf{var}Z_1 = \sigma^2 \in (0, \infty)$, $\mathsf{E}|Z_1|^3 = \rho < \infty$. We write F for the distribution function of Z_1 and F_v stands for the distribution function of $(Z_1 + \ldots + Z_v)/(\sigma\sqrt{v})$. Then (see, e.g., Theorem 5.4 of [43]), for any $v \in \mathbb{N}$,

$$\sup_{u\in\mathbb{R}} |F_v(u) - \Phi(u)| \leq \frac{C_0\rho}{\sigma^3\sqrt{v}},$$

where $\Phi(u)$ is the distribution function of a standard normal random variable, C_0 is a positive constant (C_0 does not depend on distribution of Z_1 and v). According to [44] one has $C_0 \leq 0,4693$. Consequently, taking $Z \sim N(0,1)$, we have

$$\mathsf{P}\left(\left|\frac{\sum_{q\in\overline{D}_k(N)}(\zeta_q - \mathsf{E}\zeta_q)}{\sigma_0\sqrt{M_k}}\right| > \frac{p_0^2\sqrt{M_k}\varepsilon_N}{8\sharp\mathbb{X}\sigma_0}\right) \leq \mathsf{P}\left(|Z| > \frac{p_0^2\sqrt{M_k}\varepsilon_N}{8\sharp\mathbb{X}\sigma_0}\right) + \frac{2C_0}{\sigma_0^3\sqrt{M_k}} \tag{43}$$

since $\mathsf{E}|\zeta_q - \mathsf{E}\zeta_q|^3 \leq 1$ for $q \in \overline{D}_k(N)$, where $\zeta_q = \mathbb{I}\{X_U^{(q)} = x_U, Y^{(q)} = y\}$.

It is well-known (see, e.g., formula (29) of Chap. II of [41]), that, for $u > 0$, the following inequality is true:

$$\mathsf{P}(|Z| \geq u) \leq \frac{\sqrt{2/\pi}}{u}\exp\left\{-\frac{u^2}{2}\right\}.$$

Therefore, by virtue of an inequality $\sigma_0^2 \leq 1/4$ (which is valid for the indicator variance) and as

$$(K-1)[N/K] \leq M_k \leq N, \tag{44}$$

we can write under condition (23) that

$$\mathsf{P}\left(|Z| > \frac{p_0^2\sqrt{M_k}\varepsilon_N}{8\sharp\mathbb{X}\sigma_0}\right) \leq \frac{8\sharp\mathbb{X}\sqrt{2}\sigma_0}{p_0^2\sqrt{\pi M_k}\varepsilon_N}\exp\left\{-\frac{1}{2}\left(\frac{p_0^2\sqrt{M_k}\varepsilon_N}{8\sharp\mathbb{X}\sigma_0}\right)^2\right\}$$

$$\leq \frac{4\sqrt{2}\sharp\mathbb{X}}{p_0^2\sqrt{\pi M_k}\varepsilon_N}\exp\left\{-\frac{1}{32}\left(\frac{p_0^2\sqrt{M_k}\varepsilon_N}{\sharp\mathbb{X}}\right)^2\right\}$$

$$= \frac{4\sqrt{2}\sharp\mathbb{X}}{p_0^2\sqrt{\pi M_k}}\exp\left\{-\frac{1}{32}\left(\frac{p_0^2\sqrt{M_k}\varepsilon_N}{\sharp\mathbb{X}}\right)^2+\log\left(\frac{1}{\varepsilon_N}\right)\right\}\leq\frac{C_1}{\sqrt{N}},\quad N\in\mathbb{N},$$

and C_1 does not depend on N.

Introduce

$$\widetilde{\sigma}^2:=\min_{U\subset T,x\in\mathbb{X},y\in\mathbb{Y}}\sigma_0^2(U,x,y),$$

where one considers only strictly positive $\sigma_0^2(U,x,y)$. Then obviously $\widetilde{\sigma}^2>0$, as there exists only a finite collection of different variants. Thus in view of (44), for all x, y and U under consideration, one has

$$\frac{2C_0}{\widetilde{\sigma}^3\sqrt{M_k}}\leq\frac{C_2}{\sqrt{N}},\quad N\in\mathbb{N},$$

where C_0 appeared in (43) and C_2 does not depend on N.

Therefore, if condition (23) is satisfied then, for all $x\in\mathbb{X}$, $y\in\mathbb{Y}$, $k=1,\ldots,K$ and $j\in D_k(N)$, the following inequality holds:

$$\mathsf{E}\eta_{N,k}^{(j)}\leq\frac{C_3}{\sqrt{N}},\quad N\in\mathbb{N},\tag{45}$$

where C_3 does not depend on x,y,k and N. Hence, in view of (44) we come to the relation

$$\mathsf{E}\left(\sum_{j\in D_k(N)}g_{\overline{D}_k(N)}(X^{(j)},y;\{(X^{(q)},Y^{(q)}),q\in\overline{D}_k(N)\})\right)^4$$

$$\leq\left(8C(\sharp D_k(N))^2+8(\sharp D_k(N))^4\frac{C_3^4}{N^2}\right)\sum_{(x_q,y_q)),q\in\overline{D}_N(k)}\mathsf{P}((X^{(q)},Y^{(q)})=(x_q,y_q))\leq C_4N^2,$$

where C_4 does not depend on x,y,k and N. Thus according to (41), for all N large enough, we have proved the inequality

$$\mathsf{E}U_{N,1}^4\leq C_5,\tag{46}$$

where C_5 does not depend on N.

In a similar way (taking into account (42) and (45)), for $i=0,\ldots,2m-1$, $y\in\mathbb{Y}$, $k=1,\ldots,K$, and all N large enough, we get

$$\mathsf{E}\mathbb{S}_k(i,y)^8\leq C_6(\sharp D_N(k))^{-4},\tag{47}$$

where $\mathbb{S}_k(i,y)$ is introduced in (31), and C_6 does not depend on N.

We will employ an elementary result for the Bernoulli scheme. Let $U_1,U_2,\ldots,$ be a sequence of i.i.d. random variables such that $\mathsf{P}(U_1=1)=\mathsf{p}$ and $\mathsf{P}(U_1=0)=1-\mathsf{p}$, where $\mathsf{p}\in(0,1)$. Consider the following frequency estimator of a probability p:

$$\widehat{\mathsf{p}}_N:=\frac{1}{N}\sum_{j=1}^N\mathbb{I}\{U_j=1\},\quad N\in\mathbb{N}.$$

Define

$$\widehat{\psi}_N:=\begin{cases}\frac{1}{\widehat{\mathsf{p}}_N}, & \widehat{\mathsf{p}}_N\neq 0,\\ 0, & \widehat{\mathsf{p}}_N=0.\end{cases}\tag{48}$$

Lemma 2. *For the Bernoulli scheme introduced above and the estimators $\widehat{\psi}_N$ provided by formula (48), for each $t \in \mathbb{N}$, the following relation holds:*

$$\mathsf{E}\left(\widehat{\psi}_N - \frac{1}{\mathsf{p}}\right)^t = O\left(\frac{1}{N}\right), \quad N \to \infty. \tag{49}$$

More precisely, the absolute value of the function in the left-hand side of (49), for all $N \in \mathbb{N}$, admits a bound c/N where $c = c(\mathsf{p}, t)$ for $\mathsf{p} \in (0,1)$ and $t \in \mathbb{N}$.

For the sake of completeness the proof of this result is given in Appendix A.

Now we continue the proof corresponding to Step 1. For all considered k, i, y and any $N \in \mathbb{N}$, the Cauchy - Bunyakovsky - Schwarz inequality yields

$$\mathsf{E}\big((\widehat{\psi}_{N,k}(y) - \psi(y))\mathbb{S}_k(i,y)\big)^4 \le \Big(\mathsf{E}(\widehat{\psi}_{N,k}(y) - \psi(y))^8\, \mathsf{E}\mathbb{S}_k(i,y))^8\Big)^{\frac{1}{2}}.$$

Due to Lemma 2 one has $\mathsf{E}(\widehat{\psi}_{N,k}(y) - \psi(y))^8 = O\Big(\frac{1}{N}\Big)$, $N \to \infty$. Employing the Minkowski inequality (to take into account the summation over i, y, k), for all $N \in \mathbb{N}$, we come to the bound

$$\mathsf{E}U_{N,2}^4 \le N^2 C_7 \left(\left(\frac{1}{N}\right)\left(\frac{1}{N^4}\right)\right)^{\frac{1}{2}} = \frac{C_7}{\sqrt{N}}, \tag{50}$$

where C_7 does not depend on N.

Consequently, by virtue of (33), (46) and (50) the uniform integrability of a sequence $(R_N^2)_{N\in\mathbb{N}}$ is established. Thus (28) is verified.

Step 2. Now we study the asymptotic behavior of the variables $\sqrt{N}(\widehat{T}_N(f^U) - T_N(f^U))$, as $N \to \infty$, where $\widehat{T}_N(f^U)$ and $T_N(f^U)$ are given by Formulas (26) and (27), respectively. For $j \in \mathbb{N}$, $i = 0, \dots, 2m-1$, $y \in \mathbb{Y}$, we set $Z_i^{(j)}(y) = \mathbb{I}\{Y^{(j)} = y, |f(X^{(j)}) - y| > i\}$. One has

$$\sqrt{N}(\widehat{T}_N(f^U) - T_N(f^U)) = \mathbb{W}_{N,1} + \mathbb{W}_{N,2},$$

where

$$\begin{aligned}
\mathbb{W}_{N,1} &= \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{(\widehat{\psi}_{N,k}(y) - \psi(y))}{\sharp D_k(N)}\sum_{j\in D_k(N)}(Z_i^{(j)}(y) - \mathsf{E}Z_i^{(j)}(y)),\\
\mathbb{W}_{N,2} &= \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{(\widehat{\psi}_{N,k}(y) - \psi(y))}{\sharp D_k(N)}\sum_{j\in D_k(N)}\mathsf{P}(Y^{(j)} = y, |f^U(X^{(j)}) - y| > i)\\
&= \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}(\widehat{\psi}_{N,k}(y) - \psi(y))\mathsf{P}(Y = y, |f^U(X) - y| > i).
\end{aligned} \tag{51}$$

The purpose of the second step is to prove that

$$\mathsf{E}\mathbb{W}_{N,1}^2 \to 0, \quad N \to \infty. \tag{52}$$

For $k = 1, \dots, K$, $i = 0, \dots, 2m-1$ and $y \in \mathbb{Y}$ introduce

$$\mathbb{G}_k(i,y) = \frac{1}{\sharp D_k(N)}\sum_{j\in D_k(N)}(Z_i^{(j)}(y) - \mathsf{E}Z_i^{(j)}(y)).$$

The Cauchy-Bunyakovsky-Schwarz inequality yields

$$\mathsf{E}\big((\widehat{\psi}_{N,k}(y) - \psi(y))\mathbb{G}_k(i,y)\big)^2 \le \Big(\mathsf{E}(\widehat{\psi}_{N,k}(y) - \psi(y))^4\Big)^{\frac{1}{2}}\Big(\mathsf{E}(\mathbb{G}_k(i,y))^4\Big)^{\frac{1}{2}}.$$

For each considered N, y, i and k, the variables $\{Z_i^{(j)}(y), j \in D_k(N)\}$ are independent and $|Z_i^{(j)}(y) - \mathsf{E}Z_i^{(j)}(y)| \le 1$, so by virtue of the Rosenthal inequality (42) we obtain

$$\mathsf{E}\left(\sum_{j\in D_k(N)}(Z_i^{(j)}(y) - \mathsf{E}Z_i^{(j)}(y))\right)^4 = O(\sharp D_k(N)^2).$$

Taking into account Lemma 2 for $t = 4$ and in view of (44), for each $k = 1, \dots, K$, we get the relation

$$\mathsf{E}\mathbb{W}_{N,1}^2 = O\left(N^{-\frac{1}{2}}\right), \quad N \to \infty.$$

Therefore, the goal of the second step has been achieved.

Step 3. The implementation of steps 1 and 2 permits to reduce the study of the asymptotic behavior (as $N \to \infty$) of $\mathbb{Z}_N$ given by Formula (25) to the study of variables

$$\eta_N := \sqrt{N}(T_N(f^U) - \mathsf{Err}(f^U)) + \mathbb{W}_{N,2}, \quad N \in \mathbb{N},$$

where $\mathbb{W}_{N,2}$ is defined by Formula (51).

The aim of the third step is to prove that $\mathsf{E}(\eta_N)^2 \to \sigma^2(U)$, as $N \to \infty$, where $\sigma^2(U)$ is the variance of the random variable $V(U)$ appearing in Formula (22).

On this way, we will show that the sum of certain part of the terms in a specified representation of the variables η_N does not affect (in the sense of $L^2(\Omega, \mathcal{F}, \mathsf{P})$) the limit behavior of these variables for growing N. For $y \in \mathbb{Y}$ and $W_N \subset \{1, \dots, N\}$, where $N \in \mathbb{N}$, we introduce the event

$$B_{W_N}(y) := \{\omega : \widehat{P}(y, \xi(W_N)) \ne 0\}, \tag{53}$$

where $\widehat{P}(y, \xi(W_N))$ is defined according to (13). Then, in view of the independence of observations $\xi^{(1)}, \xi^{(2)}, \dots$ we have

$$\mathsf{P}(\overline{B}_{W_N}(y)) = \mathsf{P}\left(\bigcap_{j\in W_N}\{Y^{(j)} \ne y\}\right) = (1 - \mathsf{P}(Y = y))^{\sharp W_N}.$$

If $\omega \in \overline{B}_{W_N}(y)$ then $|\widehat{\psi}(y, \xi(W_N)) - \psi(y)| = \psi(y)$. Set

$$H_N := \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}(\widehat{\psi}_{N,k}(y) - \psi(y))\mathbb{I}\{B_{N,k}(y)\}\mathsf{P}(Y = y, |f(X) - y| > i),$$

where $B_{N,k}(y) := B_{D_k(N)}(y)$ and an event $B_{W_N}(y)$ is introduced by Formula (53). Then

$$\mathsf{E}(\mathbb{W}_{N,2} - H_N)^2 = \mathsf{E}\left(\frac{\sqrt{N}}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\frac{\mathbb{I}\{\overline{B}_{N,k}(y)\}}{\mathsf{P}(Y = y)}\mathsf{P}(Y = y, |f^U(X) - y| > i)\right)^2$$

$$\le \frac{N(2m)^4}{p_0^2}\max_{y\in\mathbb{Y}}(1 - \mathsf{P}(Y = y))^{[N/K]} \to 0, \quad N \to \infty,$$

since $\sharp D_k(N) \ge [N/K]$ for $N \in \mathbb{N}$, $k = 1, \dots, K$ and because all $\mathsf{P}(Y = y) > 0$ for each $y \in \mathbb{Y}$, $[\cdot]$ stands for an integer part of a number.

We verify that H_N for large N is approximated in the space $L^2(\Omega, \mathcal{F}, \mathsf{P})$ by the random variable

$$\widetilde{H}_N := \frac{\sqrt{N}}{K}\sum_{k=1}^{K}\sum_{i=0}^{2m-1}\sum_{i-m<|y|\le m}\mathbb{I}\{B_{N,k}(y)\}\left(\frac{\mathsf{P}(Y = y) - \widehat{p}_{N,k}(y)}{\mathsf{P}(Y = y)^2}\right)\mathsf{P}(Y = y, |f^U(X) - y| > i),$$

where $\widehat{p}_{N,k}(y) := \widehat{P}(y, \xi(D_k(N)))$ and $\widehat{P}(y, \xi(W_N))$ was introduced by (13) for $y \in \mathbb{Y}$ and $W_N \subset \{1, \dots, N\}$. Evidently, $0 \le \mathsf{P}(Y = y, |f^U(X) - y| > i) \le 1$ for all k, i, y and N under consideration. Consequently, it follows that

$$\begin{aligned}
\Delta_{N,k}(i,y) &:= \left|\sqrt{N}\mathbb{I}\{B_{N,k}(y)\}\left(\frac{1}{\widehat{p}_{N,k}(y)} - \frac{1}{\mathsf{P}(Y=y)}\right)\mathsf{P}(Y=y, |f^U(X)-y| > i)\right. \\
&- \left.\sqrt{N}\mathbb{I}\{B_{N,k}(y)\}\left(\frac{\mathsf{P}(Y=y) - \widehat{p}_{N,k}(y)}{\mathsf{P}(Y=y)^2}\right)\mathsf{P}(Y=y, |f^U(X)-y| > i)\right| \\
&\le \sqrt{N}\left|\frac{\mathsf{P}(Y=y) - \widehat{p}_{N,k}(y)}{\mathsf{P}(Y=y)}\right|\left|\widehat{\psi}_{N,k}(y) - \frac{1}{\mathsf{P}(Y=y)}\right| \\
&= \frac{\sqrt{N}}{\mathsf{P}(Y=y)\sqrt{\sharp D_k(N)}}\left|\widehat{\psi}_{N,k}(y) - \frac{1}{\mathsf{P}(Y=y)}\right|\mathbb{J}_N,
\end{aligned}$$

where

$$\mathbb{J}_N := \frac{1}{\sqrt{\sharp D_k(N)}} \sum_{j \in D_k(N)} (\mathbb{I}\{Y^{(j)} = y\} - \mathsf{P}(Y^{(j)} = y)).$$

For any considered k, i, y and N the Cauchy - Bunyakovsky - Schwarz inequality implies that

$$\mathsf{E}(\Delta_{N,k}(i,y))^2 \le \frac{N}{(\mathsf{P}(Y=y)^2 \sharp D_k(N)}\left(\mathsf{E}\mathbb{J}_N^4 \mathsf{E}\left(\widehat{\psi}_{N,k}(y) - \frac{1}{\mathsf{P}(Y=y)}\right)^4\right)^{\frac{1}{2}}.$$

The Rosenthal inequality (42) yields that $\mathsf{E}\mathbb{J}_N^4 \le 2C(4)$. By means of Lemma 2 (for $t = 4$ and multipliers $c(\mathsf{p}, t)$ with $\mathsf{p} = \mathsf{P}(Y = y)$), for all considered i, y, k and any $N \in \mathbb{N}$ we come to the bound

$$\mathsf{E}(\Delta_{N,k}(i,y))^2 \le \frac{N}{(\mathsf{P}(Y=y)^2 \sharp D_k(N)} \frac{(2C(4)c(\mathsf{P}(Y=y), 4))^{\frac{1}{2}}}{\sqrt{N}}.$$

Therefore, $\mathsf{E}(H_N - \widetilde{H}_N)^2 \to 0$ as $N \to \infty$.

Let us define the variable G_N by formula similar to $\widetilde{H}_N$ but without the multiplier $\mathbb{I}\{B_{N,k}(y)\}$. In view of (44) it is easily seen that

$$\mathsf{E}(\widetilde{H}_N - G_N)^2 \le \frac{N(2m)^4}{p_0^4} \max_{y \in \mathbb{Y}} (1 - \mathsf{P}(Y=y))^{[N/K]} \left(\frac{1}{4}\right) \max_{k=1,\dots,K} \frac{1}{\sharp D_k(N)} \to 0, \quad N \to \infty.$$

Thus $\mathsf{E}(\eta_N - Q_N)^2 \to 0$ as $N \to \infty$, where

$$Q_N := \sqrt{N}(T_N(f^U) - \mathsf{Err}(f^U)) + G_N, \quad N \in \mathbb{N}.$$

Taking into account Formula (6) for the function $f = f^U$, we come to the relation

$$\begin{aligned}
Q_N &= \frac{\sqrt{N}}{K} \sum_{k=1}^{K} \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{1}{\sharp D_k(N)} \sum_{j \in D_k(N)} \left(\frac{\mathbb{I}\{Y^{(j)} = y, |f^U(X^{(j)}) - y| > i\}}{\mathsf{P}(Y=y)}\right. \\
&- \frac{\mathsf{P}(Y=y, |f^U(X)-y| > i)}{\mathsf{P}(Y=y)} + \left.\frac{(\mathsf{P}(Y=y) - \mathbb{I}\{Y^{(j)} = y\})\mathsf{P}(Y=y, |f^U(X)-y| > i)}{\mathsf{P}(Y=y)^2}\right) \\
&= \frac{\sqrt{N}}{K} \sum_{k=1}^{K} \frac{1}{\sharp D_k(N)} \sum_{j \in D_k(N)} V^{(j)},
\end{aligned} \tag{54}$$

where, for $j \in \mathbb{N}$,

$$V^{(j)} := \sum_{i=0}^{2m-1} \sum_{i-m<|y|\le m} \frac{\mathbb{I}\{Y^{(j)}=y\}}{\mathsf{P}(Y=y)} \Big(\mathbb{I}\{|f^U(X^{(j)})-y|>i\} - \mathsf{P}(|f^U(X)-y|>i|Y=y)\Big). \quad (55)$$

The variables $\{V^{(j)}, j \in \mathbb{N}\}$ are centered, i.i.d. and uniformly bounded for all j (clearly, $V^{(j)} = V^{(j)}(U)$). For each $j \in \mathbb{N}$, the distributions of $V^{(j)}$ and $V(U)$ coincide, where $V(U)$ is introduced in (22). Thus, one has

$$\mathsf{var}\, V^{(j)} = \mathsf{var}\, V(U) = \sigma^2(U), \quad j \in \mathbb{N}. \quad (56)$$

According to the lemma on grouping independent random variables, for each $N \in \mathbb{N}$, the variables $\sum_{j\in D_k(N)} V^{(j)}$, $k = 1,\dots,K$, are independent. Since $N/\sharp D_k(N) \to K$ as $N \to \infty$, for $k = 1,\dots,K$, we come to the relation

$$\mathsf{E}(Q_N^2) = \mathsf{var}\, Q_N = \frac{N}{K^2}\sum_{k=1}^{K} \frac{1}{(\sharp D_k(N))^2} \sum_{j\in D_k(N)} \mathsf{var}\, V^{(j)} = \sigma^2(U)\frac{1}{K^2}\sum_{k=1}^{K}\frac{N}{\sharp D_k(N)} \to \sigma^2(U),$$

as $N \to \infty$. Hence $\mathsf{E}\eta_N^2 \to \sigma^2(U)$, $N \to \infty$. The goal of the third step has been achieved.

In view of the above approximations (in $L^2(\Omega, \mathcal{F}, \mathsf{P})$) of the initial random variables $\mathbb{Z}_N$, introduced by (25), we conclude that $\mathsf{E}\mathbb{Z}_N^2 \to \sigma^2(U)$, as $N \to \infty$. Namely, we apply the following elementary statement: if $\mathsf{E}\alpha_N^2 \to 0$ and $\mathsf{E}\beta_N^2 \to \sigma^2$ then $\mathsf{E}(\alpha_N+\beta_N)^2 \to \sigma^2$, as $N \to \infty$. Therefore, (24) is established. The proof of Lemma 1 is complete. □

Further we will also employ a result that immediately follows from Theorem 1.

Corollary 1. *Let the conditions of Lemma 1 be satisfied. Then the following relations hold:*

$$\sqrt{N}\mathsf{E}\Big(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U) - \mathsf{Err}(f^U)\Big) \to 0, \quad N \to \infty, \quad (57)$$

$$\mathsf{var}\,(\sqrt{N}\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U)) \to \sigma^2(U), \quad N \to \infty, \quad (58)$$

where $\sigma^2(U)$ is a variance of the random variable $V(U)$ introduced in (22).

Proof. Condition (23) implies (20). Thus, according to Theorem 1, we have

$$\mathbb{Z}_N \xrightarrow{\mathcal{D}} Z \sim N(0,\sigma^2(U)), \quad N \to \infty, \quad (59)$$

where $\mathbb{Z}_N$, $N \in \mathbb{N}$, are defined in (25). Due to Lemma 1 one has the uniform integrability of the sequence $(\mathbb{Z})_{N\in\mathbb{N}}$. Consequently, relation (59) implies (57), i.e., $\mathsf{E}\mathbb{Z}_N \to \mathsf{E}Z = 0$, as $N \to \infty$. Obviously,

$$\mathsf{var}\,(\sqrt{N}\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U))$$
$$= \mathsf{E}\Big(\sqrt{N}(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U) - \mathsf{Err}(f^U)\Big)^2 - \Big(\sqrt{N}\mathsf{E}(\widehat{\mathsf{Err}}_{K,N,\varepsilon_N}(f^U) - \mathsf{Err}(f^U))\Big)^2.$$

Therefore, to obtain (58), it is sufficient to use Lemma 1 and take into account (57). The proof is complete. □

Note that (59) can be obtained directly under conditions of Lemma 1. For each $N \in \mathbb{N}$ and any $k = 1,\dots,K$, according to Lindeberg's theorem applied to arrays $\{V^{(j)}, j \in D_k(N)\}$ of centered i.i.d. uniformly bounded summands, where a sequence $(V^{(j)})_{j\in\mathbb{N}}$ is introduced in (55), taking into account (56) one has

$$V_{N,k} := \frac{1}{\sqrt{\sharp D_k(N)}} \sum_{j\in D_k(N)} V^{(j)} \xrightarrow{\mathcal{D}} Z_k \sim N(0,\sigma^2(U)), \quad N \to \infty. \quad (60)$$

For every $N \in \mathbb{N}$, the random variables $V_{N,k}$, $k = 1, \ldots, K$, are independent and $\operatorname{var} V_{N,k} = \sigma^2(U)$. Since $N/\sharp D_k(N) \to K$ as $N \to \infty$, for $k = 1, \ldots, K$, by virtue of (60) we come to relation

$$Q_N \xrightarrow{\mathcal{D}} Z \sim N(0, \sigma^2(U)), \quad N \to \infty, \tag{61}$$

where in view of (54) one has $Q_N = \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{N}{\sharp D_k(N)}} V_{N,k}$, $N \in \mathbb{N}$. Applying (61) and Slutsky's lemma, we arrive at (59).

Also note that relation (29) can be easily derived from (36) and (39) without employment of [32].

4. Forward Selection of Relevant Factors

Now we can turn to the sequential selection of factors based on MDR-EFE method. At the first step one searches for $j_1 \in T$ a point where the function $\widehat{Err}_{K,N,\varepsilon_N}(f^{\{i\}})$ attains the minimum over all $i \in T$. If there are several such points, then we take, e.g., one with the smallest index value. Recall that according to (17) (more precisely, after regularization), the random variable $\widehat{Err}_{K,N,\varepsilon_N}(f^{\{i\}})$ is in fact a function of $\widehat{f}_{PA}^{\{i\}}$, which is a forecast of the function $f^{\{i\}}$. Then this procedure is repeated, namely, if at $(k-1)$-th step the set $S_{k-1} := \{j_1, \ldots, j_{k-1}\}$ is constructed, where $k \in \{2, \ldots, r\}$, then $j_k \in T \setminus S_{k-1}$ is selected at step k in such a way that given $j_1, \ldots, j_{k-1}$ the function $\widehat{Err}_{K,N,\varepsilon_N}(f^{\{S_{k-1},i\}})$ takes the minimum value over $i \in T \setminus S_{k-1}$ for $i = j_k$. It is convenient to assume that an empty set is taken at the zero step. Then at each next step one new element is added to the previously constructed sets. If at some step there are several minimum points of the considered function then we take only one of them, e.g., with the minimal index.

Thus, for each $N \in \mathbb{N}$ the random sets $S_k(N) = S_k(N, \omega) := \{j_1, \ldots, j_k\}$ arise, where $k = 1, \ldots, r$ and $j_m = j_m(N, \omega)$, $m = 1, \ldots, r$. By construction one can write

$$j_k(N, \omega) \in J_k(N, \omega) := \operatorname{arg} \min_{i \in T \setminus S_{k-1}(N,\omega)} \widehat{Err}_{K,N,\varepsilon_N}(f^{\{S_{k-1}(N,\omega),i\}}),$$

where $S_0 := \varnothing$ and $\{\varnothing, i\} := \{i\}$. In other words the choice $j_k(N, \omega)$ at step k means that, for $i \in T \setminus S_{k-1}(N, \omega)$,

$$\widehat{Err}_{K,N,\varepsilon_N}(f^{S_k(N,\omega)}) \le \widehat{Err}_{K,N,\varepsilon_N}(f^{\{S_{k-1}(N,\omega),i\}}), \tag{62}$$

moreover, $j_k(N, \omega) = \min\{i : i \in J_k(N, \omega)\}$, $k = 1, \ldots, r$. If the joint distribution of X and Y is known, then instead of the described scheme for constructing random sets, $S_k(N, \omega)$ we turn to considering the non-random "oracle" sets $T_k = \{i_1, \ldots, i_k\}$, where $k = 1, \ldots, r$,

$$i_k \in \operatorname{arg} \min_{i \in T \setminus T_{k-1}} \mathsf{Err}(f^{\{T_{k-1},i\}}), \tag{63}$$

$T_0 := \varnothing$, and the functional Err is introduced by formula (2). If there are several i_k satisfying (63) we take among them that one which has the minimal value.

For $k \in \{1, \ldots, r\}$ and $i \in T \setminus T_k$ introduce

$$C_{k,i} := \mathsf{Err}(f^{\{T_{k-1},i\}}) - \mathsf{Err}(f^{T_k}).$$

By construction of the sets T_k we have $C_{k,i} \ge 0$, where $k = 1, \ldots, r$ and $i \in T \setminus T_k$. We call a model, satisfying condition (1), *regular* whenever the following relation is true:

$$C_{k,i} > 0, \quad k = 1, \ldots, r, \quad i \in T \setminus T_k. \tag{64}$$

In other words, for each $k = 1, \ldots, r$, a point i_k in (63) is determined uniquely. Further we employ the penalty function introduced in (11). We also use its strongly consistent estimate of type (48) with

$$\widehat{p}_N := \frac{1}{W_N} \sum_{j \in W_N} \mathbb{I}\{Y^{(j)} = y\}, \tag{65}$$

$W_N \subset \{1, \ldots, N\}$ and $\sharp W_N \to \infty$ as $N \to \infty$.

Theorem 2. *Let the considered model (1) with a collection of relevant factors having cardinality $r < p$, be regular, i.e., let (64) take place. Then, for the random sets $S_r(N)$ introduced above, the following relation is valid*

$$\mathsf{P}(S_r(N) = T_r) \to 1, \quad N \to \infty, \tag{66}$$

where T_r is defined by means of (63) for $k = 1, \ldots, r$. In other words, with probability close to one, the described procedure of forward selection based on statistical estimates of the error functional leads to the "oracle" collection T_r, when N is large enough.

Proof. For a random set $S_r(N, \omega) = \{j_1(N, \omega), \ldots, j_r(N, \omega)\}$, where $j_k(N, \omega)$ is an element taken at k-th step, one has

$$\mathsf{P}(\omega : S_r(N, \omega) = T_r) \geq \mathsf{P}(\omega : j_1(N, \omega) = i_1, \ldots, j_r(N, \omega) = i_r).$$

Note that

$$\mathsf{P}(\omega : j_1(N, \omega) = i_1, \ldots, j_r(N, \omega) = i_r) \geq \mathsf{P}\left(\bigcap_{k=1}^{r} A_k(N)\right),$$

where

$$A_k(N) := \bigcap_{i \in T \setminus T_{k-1}} \left\{\widehat{Err}_{K,N,\varepsilon_N}(f^{T_k}) < \widehat{Err}_{K,N,\varepsilon_N}(f^{\{T_{k-1}, i\}})\right\},$$

$k = 1, \ldots, r$. Thus, we obtain:

$$\mathsf{P}\left(\bigcap_{k=1}^{r} A_k(N)\right) = 1 - \mathsf{P}\left(\bigcup_{k=1}^{r} \overline{A}_k(N)\right) \geq 1 - \sum_{k=1}^{r} \mathsf{P}\left(\overline{A}_k(N)\right)$$

$$\geq 1 - \sum_{k=1}^{r} \sum_{i \in T \setminus T_{k-1}} \mathsf{P}\left(\widehat{Err}_{K,N,\varepsilon_N}(f^{T_k}) \geq \widehat{Err}_{K,N,\varepsilon_N}(f^{\{T_{k-1}, i\}})\right), \tag{67}$$

where, as usual, $\overline{A} := \Omega \setminus A$ for $A \subset \Omega$. Then, for $k = 1, \ldots, r$, $i \in T \setminus T_{k-1}$ and $N \in \mathbb{N}$, we get

$$\begin{aligned} \Delta_{k,i}(N) &:= \widehat{Err}_{K,N,\varepsilon_N}(f^{T_k}) - \widehat{Err}_{K,N,\varepsilon_N}(f^{\{T_{k-1}, i\}}) \\ &= (\widehat{Err}_{K,N,\varepsilon_N}(f^{T_k}) - \mathsf{E}\widehat{Err}_{K,N,\varepsilon_N}(f^{T_k})) + (\mathsf{E}\widehat{Err}_{K,N,\varepsilon_N}(f^{T_k}) - \mathsf{Err}(f^{T_k})) \\ &+ (\mathsf{Err}(f^{T_k}) - \mathsf{Err}(f^{\{T_{k-1}, i\}})) + (\mathsf{Err}(f^{\{T_{k-1}, i\}}) - \mathsf{E}\widehat{Err}_{K,N,\varepsilon_N}(f^{\{T_{k-1}, i\}})) \\ &+ (\mathsf{E}\widehat{Err}_{K,N,\varepsilon_N}(f^{\{T_{k-1}, i\}}) - \widehat{Err}_{K,N,\varepsilon_N}(f^{\{T_{k-1}, i\}})). \end{aligned} \tag{68}$$

For $U \subset T$, set

$$Z_N(U) := \widehat{Err}_{K,N,\varepsilon_N}(f^U) - \mathsf{E}\widehat{Err}_{K,N,\varepsilon_N}(f^U).$$

For any $k = 1, \ldots, K$, $i \in T \setminus T_{k-1}$ and each $\delta \in (0, 1)$ in light of formula (57) of Corollary 1, for all N large enough ($N \geq N_2(\delta, k, i)$) it holds

$$\mathsf{P}(\Delta_{k,i}(N) \geq 0) \leq \mathsf{P}(\sqrt{N}|Z_N(T_k(N))| + \sqrt{N}|Z_N(\{T_{k-1}(N), i\})| \geq \sqrt{N}C_{k,i} - \delta)$$

$$\leq \mathsf{P}\left(\sqrt{N}|Z_N(T_k(N))| \geq \frac{(1-\delta)\sqrt{N}C_{k,i}}{2}\right) + \mathsf{P}\left(|Z_N(\{T_{k-1}(N), i\})| \geq \frac{(1-\delta)\sqrt{N}C_{k,i}}{2}\right),$$

where $C_{k,i}$ are introduced in (66), $\Delta_{k,i}(N)$ is defined by (68).

Applying the Bienaymé - Chebyshev inequality and taking into account Formula (58) of Corollary 1, for each $U \subset T$ and any $c > 0$, we come, for a centered random variable $Z_N(U)$, to the relation

$$\mathsf{P}(\sqrt{N}|Z_N(U)| \geq c\sqrt{N}) \leq \frac{N\mathsf{var}\, Z_N(U)}{Nc^2} \sim \frac{\mathsf{var}\, V(U)}{Nc^2}, \quad N \to \infty, \tag{69}$$

where $V(U)$ is determined by Formula (22). According to (64), for $k \in \{1, \ldots, r\}$ and $i \in T \setminus T_k$, one has $C_{k,i} > 0$. Therefore, for all N large enough ($N \geq N_3(\delta, k, i)$), the following inequality takes place:

$$\mathsf{P}(\Delta_{k,i}(N) \geq 0) \leq \frac{4(\mathsf{var}\, V(T_k) + \mathsf{var}\, V(\{T_{k-1}, i\})}{N(1-\delta)^2 C_{k,i}^2}. \tag{70}$$

For a fixed $m \in \mathbb{N}$, one can change the summation order over i and y to write Formula (22) as follows:

$$V(U) = \sum_{y=-m}^{m} \frac{\mathbb{I}\{Y = y\}}{\mathsf{P}(Y = y)} W(y, U),$$

where

$$W(y, U) = \sum_{0 \leq i < |y|+m} \Big(\mathbb{I}\{|f^U(X) - y| > i\} - \mathsf{P}(|f^U(X) - y| > i | Y = y)\Big). \tag{71}$$

Thus, for any $U \subset T$, one has

$$|V(U)| \leq 2m \sum_{y=-m}^{m} \frac{\mathbb{I}\{Y = y\}}{\mathsf{P}(Y = y)}.$$

Consequently, we come to the inequality

$$\mathsf{var}\, V(U) \leq \mathsf{E}V^2(U) \leq 4m^2 \sum_{y=-m}^{m} \frac{1}{\mathsf{P}(Y = y)} =: a,$$

where $a = a(m, (\mathsf{P}(Y = y))_{y \in \mathbb{Y}})$. We see that $\mathsf{var}\, V(T_k) + \mathsf{var}\, V(\{T_{k-1}, i\}) \leq 2a$ for all $k \in \{1, \ldots, r\}$, $i \in T \setminus T_{k-1}$ and $N \in \mathbb{N}$. For each $\delta \in (0, 1)$, any $k \in \{1, \ldots, r\}$, $i \in T \setminus T_{k-1}$ and all N large enough, we get the following bound:

$$\mathsf{P}(\Delta_{k,i}(N) \geq 0) \leq \frac{8a}{N(1-\delta)^2 C_{k,i}^2}.$$

Hence, for each $\delta \in (0, 1)$ and all N large enough, by virtue of (67) the following inequality holds:

$$\mathsf{P}(S_r(N) = T_r) \geq 1 - \frac{8ar}{N(1-\delta)^2 C_0^2}\left(p + 1 - \frac{r+1}{2}\right), \tag{72}$$

where $C_0^2 := \min_{k=1,\ldots,r,\, i \in T \setminus T_{k-1}} C_{k,i}^2 > 0$ according to (64). Thus relation (72) implies the validity of (66). □

Now note that according to (69) the following relation is true:

$$\mathsf{P}(\sqrt{N}|Z_N(U)| \geq c\sqrt{N}) = O\left(\frac{1}{N}\right), \quad N \to \infty. \tag{73}$$

The question arises whether this probability decreases like C/N where C is a positive constant or more rapidly. The answer depends on the variance of the random variable $V(U)$ given by Formula (22). In view of (70) we will determine when the variable $V(U)$ is degenerate, i.e., equal to a constant a.s. This is also of independent interest for the CLT established in Section 6 of [32] and given above as Theorem 1. The following result provides a simple characterization of the $V(U)$ degeneracy.

Lemma 3. *For an arbitrary set $U \subset T$, the variance of the random variable $V(U)$, appearing in Formula (22), is zero if and only if, for every $y \in \mathbb{Y}$, there is $k_0(y) \in \{0, \dots, m + |y|\}$ such that*

$$\mathsf{P}(|f^U(X) - y| = k_0(y), Y = y) = \mathsf{P}(Y = y). \tag{74}$$

Thus, for each $y \in \mathbb{Y}$, on the set $\{Y = y\}$ the random variable $f^U(X)$ does not necessarily take a constant value. Moreover, the values of $k_0(y)$ need not coincide for different y.

Proof. For $y = 0, \dots, m$ and a random variable $W(y, U)$, introduced by Formula (71), one can write

$$\begin{aligned}
W(y,U) &= \sum_{0 \leq i < y+m} \Big(\mathbb{I}\{|f^U(X) - y| > i\} - \mathsf{P}(|f^U(X) - y| > i | Y = y) \Big) \\
&= \sum_{0 \leq i < y+m} \sum_{i < k \leq m+y} \Big(\mathbb{I}\{|f^U(X) - y| = k\} - \mathsf{P}(|f^U(X) - y| = k | Y = y) \Big) \\
&= \sum_{k=1}^{m+y} \sum_{i=0}^{k-1} \Big(\mathbb{I}\{|f^U(X) - y| = k\} - \mathsf{P}(|f^U(X) - y| = k | Y = y) \Big) \\
&= \sum_{k=1}^{m+y} k(\mathbb{I}\{|f^U(X) - y| = k\} - \mathsf{P}(|f^U(X) - y| = k | Y = y)) \\
&= \sum_{k=1}^{m+y} k\mathbb{I}\{|f^U(X) - y| = k\} - \mathsf{E}(|f^U(X) - y| | Y = y).
\end{aligned}$$

In a similar way we consider $y = -m, \dots, -1$. Thus, for all $y \in \mathbb{Y}$, one gets

$$W(y,U) = \sum_{k=1}^{m+|y|} k\mathbb{I}\{|f^U(X) - y| = k\} - \mathsf{E}(|f^U(X) - y| | Y = y).$$

Recall that $\mathsf{P}(Y = y) > 0$ for all $y \in \mathbb{Y}$. If, for some $y, k, j \in \mathbb{Y}$, $k \neq j$, we have

$$\mathsf{P}(|f^U(X) - y| = k, Y = y) > 0, \;\; \mathsf{P}(|f^U(X) - y| = j, Y = y) > 0,$$

then on the events $\{|f^U(X) - y| = k, Y = y\}$ and $\{|f^U(X) - y| = j, Y = y\}$ the variable $W(y, U)$ takes different values. Therefore, $V(U)$ takes different values on these events. Hence $\mathsf{var}\, V(U) > 0$, if (74) is not valid. Thus (74) is a necessary condition to guarantee that $\mathsf{var}\, V(U) = 0$. Suppose now that, (74) holds. In this case we get

$$\mathsf{E}(|f^U(X) - y| | Y = y) = k_0(y), \;\; y \in \mathbb{Y}.$$

Clearly, $k_0(y)$ depends on U as well. We see that $V(U)$ on each set $\{Y = y\}$ takes (up to the set of measure zero) the value $\frac{1}{\mathsf{P}(Y=y)}(k_0(y) - k_0(y)) = 0$, $y \in \mathbb{Y}$. Therefore, $\mathsf{var}\, V(U) = 0$. Note that $k_0(y)$ need not coincide for different $y \in \mathbb{Y}$. The proof is complete. □

5. Concluding Remarks

The established asymptotical result (Theorem 2) is rather qualitative in nature, since relation (66) assumes increasing values of N. Relation (72) is more precise. However, (72) demonstrates that, loosely speaking, one has to employ $N >> rp$. As previously, we assume that assumption (A), introduced on page 2, is valid. Evidently, the sequential choice of relevant variables based on statistical estimators of the error functional (of response approximation), is attractive for implementation, although suboptimal. In this regard Theorem 2 shows that under certain conditions, forward (random) selection with a high probability leads to the same collection of factors, which is provided by the sequential procedure with known joint distribution of the vector of factors X and the response Y. In the future work, it would be reasonable to supplement the theoretical results by computer simulations (see, e.g., [45]).

Consideration of the proximity of the results of optimal and suboptimal procedures requires a separate study. In addition, we note that within the framework of linear models, estimates of the probability of correct identification of relevant factors are considered, e.g., in [46,47]. Theorem 2 does not assume the linearity of stochastic model. Presumably for the first time, in our work a forward selection of relevant factors affecting the non-binary random response is treated on the base of MDR-EFE method. It would be interesting to extend the conditions allowing to establish relation (66). Moreover, stability problems of FS deserve special attention, see, e.g., [48–50]. Algorithms stability for classification problems in the framework of random trees is treated in [51].

Finally, we emphasize that the problem of statistical estimation of the cardinality of a set of relevant factors appearing in definition (1) is very important and complex. Along with dealing with the deterministic number of selected factors, there is a research approach based on developing the rules for stopping the procedures used to identify the relevant set. In this regard, we indicate, e.g., article [52], dedicated to information methods for selecting relevant factors. The study of non-discrete stochastic models is also of undoubted interest, see, e.g., [53].

Further it would be interesting to study other functionals than (2) to measure the quality of a response approximation by means of functions defined on various collections of factors. One can also consider a random number of observations. In this regard we refer, e.g., to [27,54].

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The author is very grateful to the Reviewers for careful reading the manuscript and making valuable remarks and suggestions. He would also like to thank Alexander Tikhomirov for invitation to present manuscript for this issue.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Lemma 2

Proof. For any $t \in \mathbb{N}$ and $\mathsf{p} \in (0,1)$, one has

$$\begin{aligned}
\mathsf{E}(\widehat{\psi}_N)^t &= N^t \sum_{j=1}^{N} \frac{1}{j^t} \binom{N}{j} \mathsf{p}^j (1-\mathsf{p})^{N-j} \\
&= \frac{N^t}{\mathsf{p}^t (N+1)\dots(N+t)} \sum_{j=1}^{N} \frac{(j+1)\dots(j+t)}{j^t} \binom{N+t}{j+t} \mathsf{p}^{j+t} (1-\mathsf{p})^{(N+t)-(j+t)} \\
&= \frac{1}{\mathsf{p}^t} (1+h_t(N)) \sum_{i=t+1}^{N+t} \left(1 + \frac{a_1}{i-t} + \dots + \frac{a_t}{(i-t)^t}\right) \binom{N+t}{i} \mathsf{p}^i (1-\mathsf{p})^{N+t-i},
\end{aligned}$$

where $h_t(N) = O(1/N)$, as $N \to \infty$, and $a_1, \dots, a_t \in \mathbb{N}$. We do not use the explicit formulas $a_1 = t(t+1)/2, \dots, a_t = t!$. Note that

$$\sum_{i=t+1}^{N+t} \binom{N+t}{i} \mathsf{p}^i (1-\mathsf{p})^{N+t-i} = 1 - \sum_{i=0}^{t} \binom{N+t}{i} \mathsf{p}^i (1-\mathsf{p})^{N+t-i} = 1 - g_t(N),$$

where $g_t(N) := \sum_{i=0}^{t} g_{t,i}(N)$ and, for $i = t+1, \dots, N+t$, one has

$$0 \le g_{t,i}(N) := \binom{N+t}{i} \mathsf{p}^i (1-\mathsf{p})^{N+t-i} \le (N+t)^t (1-\mathsf{p})^N = O(1/N), \quad N \to \infty.$$

For each $k = 1, \dots, t$, introduce

$$q_{t,k}(N) := \sum_{i=t+1}^{N+t} \frac{1}{(i-t)^k} \binom{N+t}{i} \mathsf{p}^i (1-\mathsf{p})^{N+t-i}$$

$$= \frac{1}{\mathsf{p}^k (N+t+1)\dots(N+t+k)} \sum_{i=t+1}^{N+t} \frac{(i+1)\dots(i+k)}{(i-t)^k} \binom{N+t+k}{i+k} \mathsf{p}^{i+k} (1-\mathsf{p})^{(N+t+k)-(i+k)}.$$

Obviously, one can write $q_{t,k}(N) = O\left(1/N^k\right)$, as

$$(i+1)\dots(i+k)(i-t)^{-k} \le (1+t+k)^k \le (1+2t)^t$$

for all $i \ge t+1$, $k = 1, \dots, t$, and since

$$\sum_{i=t+1}^{N+t} \binom{N+t+k}{i+k} \mathsf{p}^{i+k} (1-\mathsf{p})^{(N+t+k)-(i+k)} \le 1.$$

Consequently, for any $t \in \mathbb{N}$, we get

$$\mathsf{E}(\widehat{\psi}_N)^t = \frac{1}{\mathsf{p}^t}(1+h_t(N)) \left(1 - g_t(N) + \sum_{k=1}^{t} q_{t,k}(N)\right) = \frac{1}{\mathsf{p}^t} + R_t(N),$$

where $R_t(N) = O(1/N)$, as $N \to \infty$. Evidently, $\mathsf{E}(\widehat{\psi}_N)^0 = 1$ for $N \in \mathbb{N}$. For each $N \in \mathbb{N}$, set $R_0(N) = 0$. Thus, for $t \in \mathbb{N}$, one has

$$\mathsf{E}\left(\widehat{\psi}_N - \frac{1}{\mathsf{p}}\right)^t = \sum_{v=0}^{t} \binom{t}{v} \mathsf{E}(\widehat{\psi}_N)^v \left(-\frac{1}{\mathsf{p}}\right)^{t-v} = \sum_{v=0}^{t} \binom{t}{v} \left(\frac{1}{\mathsf{p}^v} + R_v(N))\right) \left(-\frac{1}{\mathsf{p}}\right)^{t-v} = O\left(\frac{1}{N}\right),$$

because

$$\sum_{v=0}^{t} \binom{t}{v} \left(\frac{1}{\mathsf{p}}\right)^v \left(-\frac{1}{\mathsf{p}}\right)^{t-v} = \left(\frac{1}{\mathsf{p}} - \frac{1}{\mathsf{p}}\right)^t = 0, \quad \sum_{v=0}^{t} \binom{t}{v} \left(\frac{1}{\mathsf{p}}\right)^{t-v} = \left(1 + \frac{1}{\mathsf{p}}\right)^t$$

and

$$\max_{v=0,\dots,t} |R_v(N)| = O(1/N), \quad N \to \infty.$$

The proof of Lemma 2 is complete. □

References

1. Seber, G.A.F.; Lee, A.J. *Linear Regression Analysis*, 2nd ed.; J.Wiley and Sons Publication: Hoboken, NJ, USA, 2003.
2. Györfi, L.; Kohler, M.; Krzyżak, A.; Walk H. *A Distribution-Free Theory of Nonparametric Regression*; Springer: New York, NY, USA, 2002.
3. Matloff, N. *Statistical Regression and Classification. From Linear Models to Machine Learning*; CRC Press: Boca Raton, FL, USA, 2017.
4. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]

5. Hastie, T.; Tibshirani, R.; Wainwrigth, R. *Statistical Learning with Sparsity. The Lasso and Generalizations*; CRC Press: Boca Raton, FL, USA, 2015.
6. Bolón-Candedo, V.; Alonso-Betanzos, A. *Recent Advances in Ensembles for Feature Selection*; Springer: Cham, Switzerland, 2018.
7. Giraud, C. *Introduction to High-Dimensional Statistics*; CRC Press: Boca Raton, FL, USA, 2015.
8. Stańczyk, U.; Zielosko, B.; Jain, L.C. (Eds.) *Advances in Feature Selection for Data and Pattern Recognition*; Springer International Publishing AG: Cham, Switzerland, 2018.
9. Kuhn, M.; Johnson, K. *Feature Engineering and Selection. A Practical Approach for Predictive Models*; CRC Press: Boca Raton, FL, USA, 2020.
10. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
11. Jia, W.; Sun, M.; Lian, J.; Hou, S. Feature dimensionality reduction: A review. *Complex Intell. Syst.* **2022**, *8*, 2663–2693. [CrossRef]
12. Lyu, Y.; Feng, Y.; Sakurai, K. A survey on feature selection techniques based on filtering methods for cyber attack detection. *Information* **2023**, *14*, 191. [CrossRef]
13. Pradip, D.; Chandrashekhar, A. A comprehensive survey on feature selection in the various fields of machine learning. *Appl. Intell.* **2023**, *52*, 4543–4581.
14. Htun, H.H.; Biehl, M.; Petkov, N. Survey of feature selection and extraction techniques for stock market prediction. *Financ. Innov.* **2023**, *9*, 26. [CrossRef] [PubMed]
15. Laborda, J.; Ryoo, S. Feature Selection in a Credit Scoring Model. *Mathematics* **2021**, *9*, 746. [CrossRef]
16. Emily, M. A survey of statistical methods for gene-gene interaction in case-control genomewide association studies. *J. Société Fr. Stat.* **2018**, *159*, 27–67.
17. Tsunoda, T.; Tanaka, T.; Nakamura, Y. (Eds.). *Genome-Wide Association Studies*; Springer: Singapore, 2019.
18. Luque-Rodriguez, M.; Molina-Baena, J.; Jimenez-Vilchez, A.; Arauzo-Azofra A. Initialization of feature selection search for classification. *J. Artif. Intell. Res.* **2022**, *75*, 953–998. [CrossRef]
19. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2022**, *2*, 927312. [CrossRef]
20. Coelho, F.; Braga, A.P.; Verleysen, M.A. Mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. *Int. J. Comput. Intell. Syst.* **2016**, *9*, 726–733. [CrossRef]
21. Kozhevin, A.A. Feature selection based on statistical estimation of mutual information *Sib. Elektron. Mat. Izv.* **2021**, *18*, 720–728. [CrossRef]
22. Latt, K.Z.; Honda, K.; Thiri, M.; Hitomi, Y.; Omae, Y.; Sawai, H.; Kawai, Y.; Teraguchi, S.; Ueno, K.; Nagasaki, M.; et al. Identification of a two-SNP PLA2R1 haplotype and HLA-DRB1 alleles as primary risk associations in idiopathic membranous nephropathy. *Sci. Rep.* **2018**, *8*, 15576. [CrossRef] [PubMed]
23. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Applic* **2014**, *24*, 175–186. [CrossRef]
24. AlNuaimi, N.; Masud, M.M.; Serhani, M.A.; Zaki, N. Streaming feature selection algorithms for big data: A survey *Appl. Comput. Inform.* **2022**, *18*, 113–135. [CrossRef]
25. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Human Genet.* **2001**, *69*, 138–147. [CrossRef]
26. Gola, D.; John, J.M.M.; van Steen, K.; Kónig, I.R. A roadmap to multifactor dimensionality reduction methods. *Briefings Bioinform.* **2016**, *17*, 293–308. [CrossRef]
27. Bulinski, A.; Kozhevin, A. New version of the MDR method for stratified samples. *Stat. Optim. Inf. Comput.* **2017**, *5*, 1–18. [CrossRef]
28. Abegaz, F.; van Lishout, F.; Mahachie, J.J.M.; Chiachoompu, K.; Bhardwaj A.; Duroux, D.; Gusareva, R.S.; Wei, Z.; Hakonarson, H.; Van Steen K. Performance of model-based multifactor dimensionality reduction methods for epistasis detection by controlling population structure. *BioData Min.* **2021**, *14*, 16. [CrossRef]
29. Yang, C.H.; Hou M.F.; Chuang L.Y.; Yang C.S.; Lin Y.D. Dimensionality reduction approach for many-objective epistasis analysis *Briefings Bioinform.* **2023**, *24*, bbac512. [CrossRef]
30. Bulinski, A.; Butkovsky, O.; Sadovnichy, V.; Shashkin, A.; Yaskov, P.; Balatskiy, A.; Samokhodskaya, L.; Tkachuk, V. Statistical Methods of SNP Data Analysis and Applications. *Open J. Stat.* **2012**, *2*, 73–87. [CrossRef]
31. Bulinski, A. On foundation of the dimensionality reduction method for explanatory variables. *J. Math. Sci.* **2014**, *199*, 113–122. [CrossRef]
32. Bulinski, A.V.; Rakitko, A.S. MDR method for nonbinary response variable. *J. Multivar. Anal.* **2015**, *135*, 25–42. [CrossRef]
33. Macedo, F.; Oliveira, M.R.; Pacheco, A.; Valadas, R. Theoretical Foundations of Forward Feature Selection Methods based on Mutual Information. *Neurocomputing* **2019**, *325*, 67–89. [CrossRef]
34. Bulinski, A.V. On relevant feature selection based on information theory. *Theory Probab. Its Appl.* **2023**, *68*, 392–410. [CrossRef]
35. Rakitko, A. MDR-EFE method with forward selection. In Proceedings of the The 5th International Conference on Stochastic Methods (ICSM-5), Moscow, Russia, 23–27 November 2020. . [CrossRef]

36. Velez, D.R.; White, B.C.; Motsinger, A.A.; Bush, W.S.; Ritchie, M.D.; Williams, S.M.; Moore, J.H. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.* **2007**, *31*, 306–315. [CrossRef] [PubMed]
37. Hu, T.-C.; Moricz, F.; Taylor, R. Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Math. Hung.* **1989**, *54*, 153–162. [CrossRef]
38. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
39. Billingsley, P. *Convergence of Probability Measures*; John Wiley and Sons: New York, NY, USA, 1968.
40. Borkar, V.S. *Probability Theory: An Advanced Course*; Springer: New York, NY, USA, 1995.
41. Bulinski, A.V.; Shiryaev, A.N. *Theory of Stochastic Processes*, 2nd ed.; Fizmatlit: Moscow, Russia, 2005. (In Russian)
42. Kallenberg, O. *Foundations of Modern Probability*; Springer: New York, NY, USA, 1997.
43. Petrov, V.V. *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*; Clarendon Press: Oxford, UK, 1995.
44. Shevtsova I.G. On absolute constants in the Berry-Esseen inequality and its structural and non-uniform refinements. *Informatics Its Appl.* **2013**, *7*, 124–125.
45. Bulinski, A.V.; Rakitko, A.S. Simulation and analytical approach to the identification of significant factors. *Commun. Stat.-Simul. Comput.* **2016**, *45*, 1430–1450. [CrossRef]
46. Shah, R.D.; Samworth, R.J. Variable selection with error control: another look at stablity selection *J.R. Statist. Soc. B.* **2012**, 74, 1-26. [CrossRef]
47. Beinrucker, A.; Dogan, U.; Blanchard, G. Extensions of stability selection using subsamples of observations and covariates. *Stat. Comput.* **2016**, *26*, 1059–1077. [CrossRef]
48. Nogueira, S.; Sechidis, K.; Brown, G. On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **2018**, 18, 1-54.
49. Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1060–1073. [CrossRef]
50. Bulinski, A. Stability properties of feature selection measures. *Theory Probab. Appl.* **2024**, *69*, 3–15.
51. Bénard, C.; Biau, G.; Da Veiga, S.; Scornet, E. SIRUS: Stable and Interpretable RUle Set for classification. *Electron. J. Statist.* **2021**, *15*, 427–505. [CrossRef]
52. Mielniczuk, J. Information theoretic methods for variable selection—A review. *Entropy* **2022**, *24*, 1079. [CrossRef]
53. Linke, Y.; Borisov, I.; Ruzankin, P.; Kutsenko, V.; Yarovaya, E.; Shalnova, S. Universal Local Linear Kernel Estimators in Nonparametric Regression. *Mathematics* **2022**, *10*, 2693. [CrossRef]
54. Rachev, S.T.; Klebanov, L.B.; Stoyanov, S.V.; Fabozzi, F.J. *The Methods of Distances in the Theory of Probability and Statistics*; Springer: New York, NY, USA, 2013.