An Improved Combined Classification Rule

Majid Mojirsheibani[1], School of Mathematics & Statistics, Carleton University, Ottawa, Ontario, K1S 5B6 Canada. E-mail: majidm@math.carleton.ca

Abstract

We propose a data-based procedure for combining a number of individual classifiers in order to construct more effective classification rules. Under some regularity conditions, the resulting combined classifier turns out to be almost surely superior to each of the individual classifiers. Here, superiority means lower misclassification error rate.

1. Introduction

Consider the following standard K-class classification problem. Let $(\mathbf{X}, Y)$ be a random pair in $\Re^d \times \{0, 1, \cdots, K-1\}$. The $\Re^d$–valued vector $\mathbf{X}$ is called the feature or predictor vector, which is always observable, and Y is called the class membership and takes values in $\{0, 1, \cdots, K-1\}$. The problem of classification is to predict Y based on $\mathbf{X}$, as accurately as possible, where accuracy means lower misclassification error rate. More specifically, let g be a classifier, i.e., any map of the form $g : \Re^d \to \{0, 1, \cdots, K-1\}$, which one uses to predict Y based on $\mathbf{X}$. The problem is then to choose g in such a way that its misclassification error rate, $\mathrm{err}(g)$, defined by

$$\mathrm{err}(g) = \mathbf{P}\{g(\mathbf{X}) \neq Y\},$$

is as small as possible. The best classifier g_{B}, called the Bayes classifier, is the one with the lowest error rate; i.e., g_{B} satisfies:

$$\mathbf{P}\{g_{\mathrm{B}}(\mathbf{X}) \neq Y\} = \inf_{g:\ \Re^d \to \{0,1,\cdots,K-1\}} \mathbf{P}\{g(\mathbf{X}) \neq Y\}.$$

Here the Bayes classifier g_{B} predicts Y as belonging to class k' if the posterior probability $P(Y{=}k'|\mathbf{X}{=}\mathbf{x}) \geq P(Y{=}k|\mathbf{X}{=}\mathbf{x})$ for all $k \in \{0, \cdots, K-1\}$. Ties are usually broken in favor of the class with the smallest index k. Clearly, when the distribution of $(\mathbf{X}, Y)$ is completely known, one would try to find the optimal classifier g_{B} which attains the lowest error. Unfortunately, in practice, the underlying distribution of the random pair

[1]Supported in part by a grant from NSERC Canada.

$(\mathbf{X}, Y)$ is unknown and the only information available is a training sample of size n, $\mathbf{T}_n = \{(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)\}$. Here $(\mathbf{X}_i, Y_i)$'s are independently and identically distributed $\Re^d \times \{0, 1, \cdots, K-1\}$–vlaued observations from the distribution of $(\mathbf{X}, Y)$. The goal is then to find a data-based classification rule g_n whose conditional misclassification error rate

$$\mathrm{err}_n(g_n) = \mathbf{P}\{g_n(\mathbf{X}) \neq Y \mid \mathbf{T}_n\}$$

is somehow as small as possible. A classification rule g_n is said to be consistent if

$$\mathrm{err}_n(g_n) \longrightarrow_p \mathrm{err}(g_{\mathrm{B}}), \quad \text{as} \quad n \to \infty.$$

We say g_n is strongly consistent if the convergence holds almost surely.

Some of the popular classification rules are histogram classifiers, Nearest-Neighbor (NN) rules, Fisher's linear discriminant function (LDA), and tree classifiers. Different classifiers have different properties and their performance can depend on many different factors. For instance the LDA performs well for normal populations. In general, in a given situation, it is not clear at all as to how one should choose a classifier. In this article we propose a procedure for combining a number of candidate classifiers in order to develop more effective classification rules. The proposed procedure is easy to implement and yields improvements in the overall error rate relative to the individual classifiers. The idea of combined estimation is relatively new and goes back to Breiman's [1] "Stacked Regression", and Wolpert's [12] "Stacked Generalization". More recent relevant papers are those of LeBlanc and Tibshirani [6], Mojirsheibani [8, 9, 10]. A non-technical presentation of the subject appears in chapter 11 of Schürmann [11].

2. The Proposed Method

Let $g_{n,1}(\mathbf{x}), \cdots, g_{n,M}(\mathbf{x})$ be M individual classification rules, where each $g_{n,m}$ is a map of the form

$$g_{n,m} : \ \{\Re^d \times \{0, 1, \cdots, K-1\}\}^n \times \Re^d \longrightarrow \{0, 1, \cdots, K-1\}.$$

Let $(\mathbf{X}, Y)$ be a new (future) observation where $\mathbf{X}$ is observable but not the class membership Y. Define the vectors $\widehat{\mathbf{W}}_i$ and $\widehat{\mathbf{W}}$ according to

$$\widehat{\mathbf{W}}_i \equiv \widehat{\mathbf{W}}(\mathbf{X}_i) = (g_{n,1}(\mathbf{X}_i), \cdots, g_{n,M}(\mathbf{X}_i)) \quad \text{and} \quad \widehat{\mathbf{W}} \equiv \widehat{\mathbf{W}}(\mathbf{X}) = (g_{n,1}(\mathbf{X}), \cdots, g_{n,M}(\mathbf{X})).$$

We will consider the non-iid discretized "data": $(\widehat{\mathbf{W}}_1, Y_1), \cdots, (\widehat{\mathbf{W}}_n, Y_n)$ and the discretized "future observation" $(\widehat{\mathbf{W}}, Y)$, where Y is to be predicted. One possible combined

classifier is ψ_n, where

$$\begin{aligned}\psi_n(\mathbf{x}) &\equiv \psi_n(g_{n,1}(\mathbf{x}),\cdots,g_{n,M}(\mathbf{x}))\\ &= \text{argmax}_{0\le k\le K-1}\sum_{i=1}^{n}\mathbf{I}\{Y_i=k\}\mathbf{I}\{g_{n,m}(\mathbf{X}_i)=g_{n,m}(\mathbf{x});\ m=1,\cdots,M\}\\ &= \text{argmax}_{0\le k\le K-1}\sum_{i=1}^{n}\mathbf{I}\{Y_i=k\}\mathbf{I}\{\widehat{\mathbf{W}}_i=\widehat{\mathbf{W}}(\mathbf{x})\},\end{aligned}\tag{1}$$

where $\mathbf{I}\{A\}$ is the indicator of the set A. In the case of ties we take $\psi_n(\mathbf{x})$ to be the smallest k for which (1) holds. The above procedure may be viewed as a multinomial discriminant function applied to the non-iid "data": $\{(\widehat{\mathbf{W}}_i, Y_i)\}_{i=1}^n$. The combined classifier ψ_n of (1) and its optimal properties were studied by Mojirsheibani [9]. The problem of classifying a discrete-valued covariate vector into one of K classes, based on iid observations, is not new and can be tackled by a number of different effective procedures. For instance, in addition to the usual multinomial discrimination rule, one can also consider the more flexible kernel-based approach of Aitchinson and Aitken [1], the nearest-neighbor procedure of Hills [5], and the adaptive weighted nearest-neighbor estimator of Hall [4]. For our setup, however, the situation is not quite straightforward. To appreciate this observe that the components of the non-iid pseudo-covariate vectors $\widehat{\mathbf{W}}_1,\cdots,\widehat{\mathbf{W}}_n$ are themselves classifiers that could depend on the data in some complicated ways. As a result, standard techniques (based on iid data) are no longer appropriate for proving consistency results. In fact, it will become clear in section 3 that our proposed combined classifier is a *data-dependent partitioning rule* (disguised as a modified multinomial procedure), where the partitioning (random) is induced by the individual classifiers, with the number of cells of the partition increasing exponentially fast in M.

A closer look at (1) reveals that ψ_n can be written as

$$\psi_n(\mathbf{x}) = \text{argmax}_{0\le k\le K-1}\sum_{i=1}^{n}\mathbf{I}\{Y_i=k\}\mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i,\widehat{\mathbf{W}}(\mathbf{x}))=0\},\tag{2}$$

where $\mathrm{d_H}(\mathbf{u},\mathbf{v})$ is the Hamming distance between the vectors $\mathbf{u}$ and $\mathbf{v}$ (i.e., the number of disagreements between the corresponding components of $\mathbf{u}$ and $\mathbf{v}$). The condition $\mathrm{d_H}(\widehat{\mathbf{W}}_i,\widehat{\mathbf{W}}(\mathbf{x}))=0$ appears to be too restrictive in (2) as it requires all the M classifiers to agree at both $\mathbf{X}_i$ and $\mathbf{X}$. When M is large, this condition can alter the effectiveness of the combined classifier ψ_n because the second indicator function in (2) will then become zero quite frequently. Therefore ψ_n cannot make good local decisions; this is particularly true in the case of sparse data. In practice, it makes sense to modify the term $\mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i,\widehat{\mathbf{W}}(\mathbf{x}))=$

0} in such a way that the summand in (2) has fewer zeros. One flexible criterion is to allow a small number of disagreements between some of the components of the vectors $\widehat{\mathbf{W}}_i$ and $\widehat{\mathbf{W}}(\mathbf{x})$. More specifically, let $M' < M$ be a positive integer and let $g_{n,m_1}, \cdots, g_{n,m_{M'}}$ be any M' individual classifiers. Also, let $\widehat{\mathbf{W}}_i^{M'}$ and $\widehat{\mathbf{W}}^{M'}(\mathbf{X})$ be the restrictions of $\widehat{\mathbf{W}}_i$ and $\widehat{\mathbf{W}}(\mathbf{X})$ to the set of classifiers $(g_{n,m_1}, \cdots, g_{n,m_{M'}})$, i.e.,

$$\begin{cases} \widehat{\mathbf{W}}_i^{M'} \equiv \widehat{\mathbf{W}}^{M'}(\mathbf{X}_i) = (g_{n,m_1}(\mathbf{X}_i), \cdots, g_{n,m_{M'}}(\mathbf{X}_i)) \\ \widehat{\mathbf{W}}^{M'}(\mathbf{X}) = (g_{n,m_1}(\mathbf{X}), \cdots, g_{n,m_{M'}}(\mathbf{X})). \end{cases}$$

Similarly, let $\widehat{\mathbf{W}}_i^{M-M'}$ and $\widehat{\mathbf{W}}^{M-M'}(\mathbf{X})$ be the $(M - M')$–dimensional vectors containing the remaining $M - M'$ classifiers. Then our proposed modified version of (2) is

$$\begin{aligned} \psi_n^{\text{new}}(\mathbf{x}) &\equiv \psi_n^{\text{new}}(g_{n,1}(\mathbf{x}), \cdots, g_{n,M}(\mathbf{x})) \\ &= \text{argmax}_{0 \le k \le K-1} \sum_{i=1}^{n} \mathbf{I}\{Y_i = k\}\Big(f_1(i, M, \mathbf{x}) + f_2(i, M, \mathbf{x}) \times f_3(i, M, \mathbf{x})\Big), \quad (3) \end{aligned}$$

where

$f_1(i, M, \mathbf{x}) = \mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i, \widehat{\mathbf{W}}(\mathbf{x})) = 0\}$,
$f_2(i, M, \mathbf{x}) = \mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i^{M-M'}, \widehat{\mathbf{W}}^{M-M'}(\mathbf{x})) = 0\}$, and
$f_3(i, M, \mathbf{x}) = \mathbf{I}\{1 \le \mathrm{d_H}(\widehat{\mathbf{W}}_i^{M'}, \widehat{\mathbf{W}}^{M'}(\mathbf{x})) \le L\}$,

for some small positive integer $L < M'$. In the case of ties $\psi_n^{\text{new}}(\mathbf{x})$ picks the smallest $k \in \{0, \cdots K-1\}$ for which (3) holds. Here, L may be viewed as the tuning parameter of the combined classifier ψ_n^{new}. In other words: $\psi_n^{\text{new}}(\mathbf{x}) = k$, if the number of class k data points for which all the $M - M'$ classifiers and at least $(M' - L)$ of the remaining M' classifiers predict (correctly or incorrectly) the same class, at both the data point and the new observation, is larger than for any of the other classes. The combined classifier (3) is in a sense a class-majority vote procedure, where the class with the largest number of "yes" (=1) votes is elected. Here the voters are the $\mathbf{X}_i$'s and the vote associated with $\mathbf{X}_i$ is a 1 (i.e., a "yes") if the condition $f_1 + f_2 \cdot f_3 = 1$ holds, and zero otherwise. An interesting case is the one with $L = 1$. Larger values of L are suitable only when M (as well as M' and n) are somehow quite large. In fact, the major focus of this article is on the case of $L = 1$. That is, we are allowing the components of $\widehat{\mathbf{W}}_i^{M'}$ and $\widehat{\mathbf{W}}^{M'}(\mathbf{x})$ to have at most one disagreement.

Observe that if we put $f_3(i, M, \mathbf{x}) = \mathbf{I}\{0 \le \mathrm{d_H}(\widehat{W}_i^{M'}, \widehat{W}^{M'}(\mathbf{x})) \le L\}$, then (3) can be re-written as

$$\psi_n^{\text{new}}(\mathbf{x}) = \text{argmax}_{0 \le k \le K-1} \sum_{i=1}^{n} \mathbf{I}\{Y_i = k\}\Big(f_2(i, M, \mathbf{x}) \times f_3(i, M, \mathbf{x})\Big),$$

where $f_2(i, M, \mathbf{x})$ is as before.

3. Asymptotic Performance of ψ_n^{new}.

In order to study the large sample behavior of ψ_n^{new}, we first need to state some preliminary results. For $k = 0, \cdots, K-1$, let

$P_k(\widehat{\mathbf{W}}(\mathbf{X})) = \mathbf{E}\big(\mathbf{I}\{Y = k\} \mid \widehat{\mathbf{W}}(\mathbf{X})\big)$,

and put

$$\psi^*(\widehat{\mathbf{W}}(\mathbf{x})) = \text{argmax}_{0 \le k \le K-1} P_k(\widehat{\mathbf{W}}(\mathbf{x})), \tag{4}$$

where as before $\widehat{\mathbf{W}}(\mathbf{X}) = (g_{n,1}(\mathbf{X}), \cdots, g_{n,M}(\mathbf{X}))$. Let $\hat{P}_k(\widehat{\mathbf{W}}(\mathbf{x}))$ be some data-based version of $P_k(\widehat{\mathbf{W}}(\mathbf{x}))$. Observe that we have used the training sample $\mathbf{T}_n$ twice in finding $\hat{P}_k(\widehat{\mathbf{W}}(\mathbf{x}))$: once to find $\widehat{\mathbf{W}}(\mathbf{x})$ and a second time to find $\hat{P}_k$ itself. Define ψ_n to be the combined classifier

$$\psi_n(\widehat{\mathbf{W}}(\mathbf{x})) = \text{argmax}_{0 \le k \le K-1} \hat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})). \tag{5}$$

Let ψ be any other combined classifier and define $\text{err}_n(\psi)$, $\text{err}_n(\psi_n)$, and $\text{err}_n(\psi^*)$, the error rates of ψ, ψ_n, and ψ^* respectively, by

$\text{err}_n(\psi) = \mathbf{P}\{\psi(\widehat{\mathbf{W}}(\mathbf{X})) \neq Y \mid \mathbf{T}_n\}$,

$\text{err}_n(\psi_n) = \mathbf{P}\{\psi_n(\widehat{\mathbf{W}}(\mathbf{X})) \neq Y \mid \mathbf{T}_n\}$, and

$\text{err}_n(\psi^*) = \mathbf{P}\{\psi^*(\widehat{\mathbf{W}}(\mathbf{X})) \neq Y \mid \mathbf{T}_n\}$.

The following lemma shows that ψ^*, as defined by (4), can be viewed as the counterpart of the Bayes classifier for our combined classification setup; of course in the current setup ψ^* is also data-based (random).

Lemma. *Let ψ^* be as in* (4).

(a) *For any other combined classifier ψ one has*

$$\text{err}_n(\psi) - \text{err}_n(\psi^*) \geq \text{a.s. } 0.$$

(b) *Let ψ_n be as in* (5), *then*

$$\text{err}_n(\psi_n) - \text{err}_n(\psi^*) \leq \text{a.s. } \sum_{k=0}^{K-1} \int \left| \hat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) - \mathbf{P}\{Y = k \mid \mathbf{X} = \mathbf{x}\} \right| \mu(d\mathbf{x}),$$

where μ is the probability measure of $\mathbf{X}$.

Proof. A proof of this result for the two-class problem ($K = 2$) is given in Mojirsheibani (1999b, Lemma 1); extension to the general $K > 2$ is straightforward and will not be given here.

In the rest of this article we will assume that the distribution of $\mathbf{X}$ has a compact support $\mathbf{B}^d \subset \Re^d$. Let

$$\Pi_n \equiv \Pi_n(\mathbf{B}^d, \mathbf{T}_n, M) \ := \{A_{n,1}, \cdots, A_{n,K^M}\}$$

be a random partition of $\mathbf{B}^d$ induced by the training sample $\mathbf{T}_n$, where the K^M cells of the partition are defined as follows. Let $\mathbf{u}_1, \cdots, \mathbf{u}_{K^M}$ be all of the different M–dimensional vectors in the discrete space $\{0, 1, \cdots, K-1\}^M$. Then the ith cell of the partition is

$$\begin{aligned} A_{n,i} &= \left\{\mathbf{x} \in \mathbf{B}^d \mid \widehat{\mathbf{W}}(\mathbf{x}) = \mathbf{u}_i; \quad \mathbf{u}_i \in \{0, 1, \cdots, K-1\}^M\right\} \\ &= \left\{\mathbf{x} \in \mathbf{B}^d \mid (g_{n,1}(\mathbf{x}), \cdots, g_{n,M}(\mathbf{x})) = \mathbf{u}_i; \quad \mathbf{u}_i \in \{0, 1, \cdots, K-1\}^M\right\}. \end{aligned}$$

Observe that $\bigcup_{i=1}^{K^M} A_{n,i} = \mathbf{B}^d$ and $A_{n,i} \cap A_{n,j} = \emptyset$ for $i \neq j$; thus Π_n is a genuine partition of $\mathbf{B}^d$. Let $A_n[\mathbf{x}]$ be the unique cell that contains the point $\mathbf{x}$ and let $\mathbf{u}_n[\mathbf{x}]$ be the vector in $\{0, 1, \cdots, K-1\}^M$ corresponding to the cell $A_n[\mathbf{x}]$. Let $\mathbf{T}_n^{(\text{obs})}$ be the observed value of the training set $\mathbf{T}_n$ and consider the nonrandom family of partitions Ω_n of $\mathbf{B}^d$:

$$\Omega_n = \left\{\Pi \mid \Pi = \Pi_n(\mathbf{B}^d, \mathbf{T}_n^{(\text{obs})}, M), \quad \text{for } \mathbf{T}_n^{(\text{obs})} \in \{\mathbf{B}^d \times \{0, 1, \cdots, K-1\}\}^n\right\}.$$

Also, for any family of partitions Ω of $\mathbf{B}^d$ let $\Delta(\Omega, n)$ be the nth shatter coefficient of Ω, i.e., the combinatorial quantity

$$\Delta(\Omega, n) = \max_{\mathbf{V}} \left\{\#\text{ of different sets in } \{(A_1 \cap \mathbf{V}, \cdots, A_N \cap \mathbf{V}) \mid (A_1, \cdots, A_N) \in \Omega\}\right\},$$

where $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$ with $\mathbf{v}_i \in \mathbf{B}^d$. Finally, we also need the following notation. For $j = 1, \cdots, K-1$, let

$$\mathbf{U}_{-m}^{(j)}(\mathbf{x}) = \Big(g_{n,1}(\mathbf{x}), \cdots, g_{n,m-1}(\mathbf{x}),\ [g_{n,m}(\mathbf{x}) + j] \ (\text{mod } K),\ g_{n,m+1}(\mathbf{x}), \cdots,\ g_{n,M}(\mathbf{x})\Big), \quad (6)$$

and let $A_{n,m}^{(j)}(\mathbf{x})$ be the cell

$$A_{n,m}^{(j)}(\mathbf{x}) = \left\{\mathbf{z} \in \mathbf{B}^d \mid (g_{n,1}(\mathbf{z}), \cdots, g_{n,M}(\mathbf{z})) = \mathbf{U}_{-m}^{(j)}(\mathbf{x})\right\}. \quad (7)$$

Remark A.
The cells $A_{n,m}^{(j)}(\mathbf{x})$, $j = 1, \cdots, K-1$ and $m = 1, \cdots, M$ may be viewed as the $M(K-1)$ neighboring cells of $A_n[\mathbf{x}]$, the cell containing the point $\mathbf{x}$. They are the neighbors of $A_n[\mathbf{x}]$ in the sense that $\mathbf{u}_n[\mathbf{x}]$, the $\{0, 1, \cdots, K-1\}$–valued vector corresponding to the cell $A_n[\mathbf{x}]$, differs from $\mathbf{U}_{-m}^{(j)}(\mathbf{x})$ in one position only. Note that for any cell A_n, its $M(K-1)$ neighbors are completely determined according to (7) and (6).

Our main results may be summarized as follows: If the number of individual classifiers $M = M_n$ is allowed to increase with n (at a specified rate), then under some regularity conditions the combined classifier ψ_n^{new} is at least as good as the best individual classifier, where *good* means low misclassification error rate. More precisely, let $\Pi_n^* \equiv \Pi_n^*(\mathbf{B}^d, \mathbf{T}_n, M_n - M_n') := \{A_{n,1}^*, \cdots, A_{n,K^{(M_n - M_n')}}^*\}$ be a random partition of $\mathbf{B}^d$ induced by the training sample $\mathbf{T}_n$, based on $(M_n - M_n')$ individual classifiers only. That is, $A_{n,i}^*$, the ith cell of the partition, is $A_{n,i}^* = \{\mathbf{x} \in \mathbf{B}^d \mid \widehat{\mathbf{W}}^{M_n - M_n'}(\mathbf{x}) = \mathbf{v}_i\}$, where the $\mathbf{v}_i$'s, $i = 1, \cdots, K^{M_n - M_n'}$ are the distinct $(M_n - M_n')$–dimensional vectors in $\{0, 1, \cdots, K-1\}^{(M_n - M_n')}$. In what follows we assume that Π_n (and thus Π_n^*) has connected cells (a set S is disconnected if there are nonempty open sets A and B such that $S{=}A \cup B$ and $A \cap B$ is empty, otherwise S is connected). Of course, one must impose the extra condition that the marginal distribution of each component of the random vector $\mathbf{X}$ has a density; without this condition, there is no hope of ever having a partition with connected cells. An easy-to-visualize example of a random partition with connected cells is the one where the individual classifiers are hyperplanes and K, the number of classes, is 2.

Theorem. *Consider the combined classifier* (3). *Let* $\text{diam}(S) = \sup_{\mathbf{x},\mathbf{y} \in S} ||\mathbf{x} - \mathbf{y}||$ *be the diameter of the set* S. *If all the marginal distributions of the vector* $\mathbf{X}$ *have densities, and are jointly supported on a compact and connected set* $\mathbf{B}^d \subset \Re^d$, *and if, as* $n \to \infty$,

(a) $M_n - M_n' \longrightarrow \infty$,
(b) $K^{2M_n} \log \Delta(\Omega_n, n)/n \longrightarrow 0$,
(c) *for every* $\delta > 0$, $\quad \mu\Big(\{\mathbf{x} \mid \text{diam}(A_n^*[\mathbf{x}]) \geq \delta\}\Big) \longrightarrow$a.s. 0,
(d) $h_k(\mathbf{x}) := \mathbf{P}\{Y = k \mid \mathbf{X} = \mathbf{x}\}$ *is continuous for each* $k = 0, 1, \cdots K-1$,

then the combined classifier ψ_n^{new} *is asymptotically strongly optimal in the sense that for any one of the individual classifiers* g_n *one has*

$$\limsup_{n \to \infty} \{\text{err}_n(\psi_n^{\text{new}}) - \text{err}_n(g_n)\} \leq \text{a.s. } 0.$$

Here, $\text{err}_n(g_n) = \mathbf{P}\{g_n(\mathbf{X}) \neq Y \mid \mathbf{T}_n\}$ and $\text{err}_n(\psi_n^{\text{new}}) = \mathbf{P}\{\psi_n(\widehat{\mathbf{W}}(\mathbf{X})) \neq Y \mid \mathbf{T}_n\}$.

Remark B.
Condition (a) guarantees that the number of cells in Π_n^* increases with n, and therefore the cells will shrink. Condition (b) controls the richness or cardinality of the family Ω_n. It also implies that $K^{2M_n}/n \to 0$, i.e., the number of individual classifiers cannot grow faster

than $o(\log n)$. Condition (c) is a shrinking cell condition; it requires the cells of Π_n^* to have small diameters when n (and thus M_n') is large enough. Condition (d) is technical.

Proof of the theorem.
We prove the theorem in three steps.

STEP 1. Equivalent representation of ψ_n^{new}.
Let $g_{n,m_\ell}(\mathbf{X}_i),\ \ell = 1, \cdots, M'$ be the components of $\widehat{\mathbf{W}}_i^{M'}$, and let $g_{n,m_\ell}(\mathbf{X}_i),\ \ell = M'+1, \cdots, M$ be the components of $\widehat{\mathbf{W}}_i^{M-M'}$. Similarly, $g_{n,m_\ell}(\mathbf{x})$ is a component of $\widehat{\mathbf{W}}^{M'}(\mathbf{x})$ for $\ell = 1, \cdots, M'$, and a component of $\widehat{\mathbf{W}}^{M-M'}(\mathbf{x})$ for $\ell = M'+1, \cdots, M$. Then a little effort shows that

$$
\begin{aligned}
& \mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i, \widehat{\mathbf{W}}(\mathbf{x})) = 0\} \\
& \quad + \mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i^{M-M'}, \widehat{\mathbf{W}}^{M-M'}(\mathbf{x})) = 0\} \times \mathbf{I}\{\mathrm{d_H}(\widehat{\mathbf{W}}_i^{M'}, \widehat{\mathbf{W}}^{M'}(\mathbf{x})) = 1\} \\
= \ & \mathbf{I}\{g_{n,m}(\mathbf{X}_i) = g_{n,m}(\mathbf{x});\quad m = 1, \cdots, M\} \\
& + \mathbf{I}\{g_{n,m_\ell}(\mathbf{X}_i) = g_{n,m_\ell}(\mathbf{x});\quad \ell = M'+1, \cdots, M\} \\
& \times \Big[\mathbf{I}\Big\{\{g_{n,m_\ell}(\mathbf{X}_i) = g_{n,m_\ell}(\mathbf{x});\quad \ell = 2, \cdots, M'\} \cap \{g_{n,m_1}(\mathbf{X}_i) \neq g_{n,m_1}(\mathbf{x})\}\Big\} \\
& + \mathbf{I}\Big\{\{g_{n,m_\ell}(\mathbf{X}_i) = g_{n,m_\ell}(\mathbf{x});\quad \ell = 1, 3, \cdots, M'\} \cap \{g_{n,m_2}(\mathbf{X}_i) \neq g_{n,m_2}(\mathbf{x})\}\Big\} \\
& + \cdots \\
& + \mathbf{I}\Big\{\{g_{n,m_\ell}(\mathbf{X}_i) = g_{n,m_\ell}(\mathbf{x});\quad \ell = 1, \cdots M'-1\} \cap \{g_{n,m_{M'}}(\mathbf{X}_i) \neq g_{n,m_{M'}}(\mathbf{x})\}\Big\}\Big] \\
= \ & \mathbf{I}\{g_{n,m}(\mathbf{X}_i) = g_{n,m}(\mathbf{x});\quad m = 1, \cdots, M\} \\
& + \mathbf{I}\Big\{\{g_{n,m}(\mathbf{X}_i) = g_{n,m}(\mathbf{x});\quad \text{all } m \neq m_1\} \cap \{g_{n,m_1}(\mathbf{X}_i) \neq g_{n,m_1}(\mathbf{x})\}\Big\} \\
& + \mathbf{I}\Big\{\{g_{n,m}(\mathbf{X}_i) = g_{n,m}(\mathbf{x});\quad \text{all } m \neq m_2\} \cap \{g_{n,m_2}(\mathbf{X}_i) \neq g_{n,m_2}(\mathbf{x})\}\Big\} \\
& + \cdots \\
& + \mathbf{I}\Big\{\{g_{n,m}(\mathbf{X}_i) = g_{n,m}(\mathbf{x});\quad \text{all } m \neq m_{M'}\} \cap \{g_{n,m_{M'}}(\mathbf{X}_i) \neq g_{n,m_{M'}}(\mathbf{x})\}\Big\} \\
=: \ & \mathbf{I}_0 + \mathbf{I}_1 + \mathbf{I}_2 + \cdots + \mathbf{I}_{M'}. \qquad (8)
\end{aligned}
$$

Now observe that since $\mathbf{x}$ is in the unique cell $A_n[\mathbf{x}]$ in which $(g_{n,1}(\mathbf{x}), \cdots, g_{n,M}(\mathbf{x})) = \mathbf{u}_n[\mathbf{x}]$, one has

$$\mathbf{I}_0 = \mathbf{I}\{(g_{n,1}(\mathbf{X}_i), \cdots, g_{n,M}(\mathbf{X}_i) = \mathbf{u}_n[\mathbf{x}]\} = \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}]\}.$$

As for the terms $\mathbf{I}_1, \cdots, \mathbf{I}_{M'}$ in (8), note that for $\ell = 1, \cdots, {}_M{}'$,

$$\{g_{n,m_\ell}(\mathbf{X}_i) \neq g_{n,m_\ell}(\mathbf{x})\} = \bigcup_{j=1}^{K-1} \Big\{g_{n,m_\ell}(\mathbf{X}_i) = [g_{n,m_\ell}(\mathbf{x}) + j] \ (\mathrm{mod}\ K)\Big\}$$

Let the vector $\mathbf{U}_{-m}^{(j)}(\mathbf{x})$ and the cell $A_{n,m}^{(j)}(\mathbf{x})$ be as in (6) and (7) respectively. Then, $\mathbf{I}_1$ can be written as

$$\begin{aligned}
\mathbf{I}_1 &:= \mathbf{I}\Big\{\{g_{n,m}(\mathbf{X}_i) = g_{n,m}(\mathbf{x}); \quad \text{all } m \neq m_1\} \cap \{g_{n,m_1}(\mathbf{X}_i) \neq g_{n,m_1}(\mathbf{x})\}\Big\} \\
&= \mathbf{I}\Big\{ \bigcup_{j=1}^{K-1} \{(g_{n,1}(\mathbf{X}_i), \cdots, g_{n,M}(\mathbf{X}_i)) = \mathbf{U}_{-m_1}^{(j)}(\mathbf{x})\}\Big\} \\
&= \mathbf{I}\Big\{ \bigcup_{j=1}^{K-1} \{\mathbf{X}_i \in A_{n,m_1}^{(j)}(\mathbf{x})\}\Big\} \\
&= \mathbf{I}\{\mathbf{X}_i \in \bigcup_{j=1}^{K-1} A_{n,m_1}^{(j)}(\mathbf{x})\}.
\end{aligned}$$

Similarly, $\mathbf{I}_\ell = \mathbf{I}\{\mathbf{X}_i \in \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\}, \quad \ell = 2, \cdots, M'$. Therefor the right hand side of (8) becomes

$$\begin{aligned}
\mathbf{I}_0 + \mathbf{I}_1 + \mathbf{I}_2 + \cdots + \mathbf{I}_{M'} &= \mathbf{I}\{\mathbf{X}_i \in A_n[\mathbf{x}]\} + \sum_{\ell=1}^{M'} \mathbf{I}\{\mathbf{X}_i \in \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\} \\
&= \mathbf{I}\Big\{\mathbf{X}_i \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big\}, \quad \text{(since the cells are disjoint).}
\end{aligned}$$

Putting all the above together, our proposed combined classifier (3) can be re-written according to

$$\begin{aligned}
\psi_n^{\text{new}}(\mathbf{x}) &\equiv \text{argmax}_{0\leq k\leq K-1} \sum_{i=1}^{n} \mathbf{I}\{Y_i = k\}\mathbf{I}\Big\{\mathbf{X}_i \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big\} \\
&= \text{argmax}_{0\leq k\leq K-1} \frac{\sum_{i=1}^{n} \mathbf{I}\{Y_i = k\}\mathbf{I}\Big\{\mathbf{X}_i \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big\}}{n\mu\Big(A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big)}, \qquad (9) \\
&\qquad \text{(since the denominator does not depend on } k.)
\end{aligned}$$

STEP 2. Decomposition of $\text{err}_n(\psi_n^{\text{new}}) - \text{err}_n(g_n)$

Let ψ^* be as in (4) and observe that

$$\begin{aligned}
\text{err}_n(\psi_n^{\text{new}}) - \text{err}_n(g_n) &= \text{err}_n(\psi_n^{\text{new}}) - \text{err}_n(\psi^*) + \text{err}_n(\psi^*) - \text{err}_n(g_n) \\
&\leq \text{a.s.} \quad \text{err}_n(\psi_n^{\text{new}}) - \text{err}_n(\psi^*) + 0,
\end{aligned}$$

where the inequality follows upon replacing ψ by g_n in part (a) of the Lemma. It remains to show that $\text{err}_n(\psi_n^{\text{new}}) - \text{err}_n(\psi^*) \longrightarrow \text{a.s. } 0$.

STEP 3. $\mathrm{err}_n(\psi_n^{\mathrm{new}}) - \mathrm{err}_n(\psi^*) \longrightarrow$a.s. 0.

By part (b) of the Lemma,

$$\mathrm{err}_n(\psi_n^{\mathrm{new}}) - \mathrm{err}_n(\psi^*) \leq \text{a.s. } \sum_{k=0}^{K-1} \int_{\mathbf{B}^d} \left| \widehat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) - h_k(\mathbf{x}) \right| \, \mu(d\mathbf{x}),$$

where $h_k(\mathbf{x}) = \mathbf{P}\{Y = k \mid \mathbf{X} = \mathbf{x}\}$, and $\widehat{P}_k(\widehat{\mathbf{W}}(\mathbf{x}))$ is as in (9), i.e.,

$$\widehat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) = \frac{\sum_{i=1}^n \mathbf{I}\{Y_i = k\}\mathbf{I}\Big\{\mathbf{X}_i \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big\}}{n\mu\Big(A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big)}.$$

Thus it is sufficient to show that $\int_{\mathbf{B}^d} |\widehat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) - h_k(\mathbf{x})| \, \mu(d\mathbf{x}) \rightarrow$a.s. 0, for each k. Let $\epsilon > 0$ be given. Then by condition (d) of the theorem and the compactness of $\mathbf{B}^d$, there is a continuous function $h^\epsilon : \mathbf{B}^d \rightarrow \Re$ such that

$$\sup_{\mathbf{x} \in \mathbf{B}^d} |h_k(\mathbf{x}) - h^\epsilon(\mathbf{x})| < \epsilon \qquad \text{(Stone-Weierstrass.)}$$

Put

$$h_n^\epsilon(\mathbf{x}) = \frac{\mathbf{E}\Big(h^\epsilon(\mathbf{X})\mathbf{I}\Big\{\mathbf{X} \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big\} \mid \mathbf{T}_n\Big)}{\mu\Big(A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big)},$$

and

$$\bar{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) = \frac{\mathbf{E}\Big(\mathbf{I}\{Y = k\}\mathbf{I}\Big\{\mathbf{X} \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big\} \mid \mathbf{T}_n\Big)}{\mu\Big(A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A_{n,m_\ell}^{(j)}(\mathbf{x})\Big)}.$$

Employing the arguments used in [4], first observe that

$$\begin{aligned}
\int_{\mathbf{B}^d} \left| \widehat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) - h_k(\mathbf{x}) \right| \, \mu(d\mathbf{x}) \quad \leq \quad & \int_{\mathbf{B}^d} \left| h_k(\mathbf{x}) - h^\epsilon(\mathbf{x}) \right| \, \mu(d\mathbf{x}) \\
& + \int_{\mathbf{B}^d} \left| h^\epsilon(\mathbf{x}) - h_n^\epsilon(\mathbf{x}) \right| \, \mu(d\mathbf{x}) \\
& + \int_{\mathbf{B}^d} \left| h_n^\epsilon(\mathbf{x}) - \bar{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) \right| \, \mu(d\mathbf{x}) \\
& + \int_{\mathbf{B}^d} \left| \bar{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) - \widehat{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) \right| \, \mu(d\mathbf{x}) \\
=: \quad & E_\epsilon + B_n + C_n + D_n.
\end{aligned}$$

Clearly,

$$E_\epsilon \leq \sup_{\mathbf{x} \in \mathbf{B}^d} \left| h_k(\mathbf{x}) - h^\epsilon(\mathbf{x}) \right| < \epsilon, \qquad \text{(because of the way } h^\epsilon \text{ was chosen.)}$$

Next, the term C_n can be bounded as follows. Put $N = M'(K-1)$. Then

$$\begin{aligned}
C_n &\le \int_{\mathbf{B}^d} \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi_n} \frac{1}{\mu\Big(A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^N A_{n,m_\ell}\Big)} \\
&\qquad \times \Big|\mathbf{E}\Big(h^\epsilon(\mathbf{X})\mathbf{I}\Big\{\mathbf{X} \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^N A_{n,m_\ell}\Big\} \mid \mathbf{T}_n\Big) \\
&\qquad\qquad - \mathbf{E}\Big(\mathbf{I}\{Y=k\}\mathbf{I}\Big\{\mathbf{X} \in A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^N A_{n,m_\ell}\Big\} \mid \mathbf{T}_n\Big)\Big| \ \mu(d\mathbf{x}) \\
&= \sum_{A\in\Pi_n} \int_A \text{(the above integrand)} \ \mu(d\mathbf{x}) \\
&= \sum_{A\in\Pi_n} \mu(A) \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi_n} \frac{1}{\mu\Big(A \cup \bigcup_{\ell=1}^N A_{n,m_\ell}\Big)} \\
&\qquad \times \Big|\int_{A\cup\bigcup_{\ell=1}^N A_{n,m_\ell}} h^\epsilon(\mathbf{x}) \ \mu(d\mathbf{x}) - \int_{A\cup\bigcup_{\ell=1}^N A_{n,m_\ell}} h_k(\mathbf{x}) \ \mu(d\mathbf{x})\Big| \\
&\le \sum_{A\in\Pi_n} \mu(A) \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi_n} \frac{1}{\mu\Big(A \cup \bigcup_{\ell=1}^N A_{n,m_\ell}\Big)} \int_{A\cup\bigcup_{\ell=1}^N A_{n,m_\ell}} \Big|h^\epsilon(\mathbf{x}) - h_k(\mathbf{x})\Big| \ \mu(d\mathbf{x}) \\
&\le \sum_{A\in\Pi_n} \mu(A) \sup_{\mathbf{x}\in\mathbf{B}^d} \Big|h^\epsilon(\mathbf{x}) - h_k(\mathbf{x})\Big| \\
&\le \sum_{A\in\Pi_n} \mu(A) \cdot \epsilon \\
&= \epsilon.
\end{aligned}$$

To deal with the term D_n let ν be the probability measure of the $\mathbf{B}^d \times \{0,1\}$–valued random vector $(\mathbf{X}, \mathbf{I}\{Y=k\})$, and let ν_n be the corresponding empirical measure of $(\mathbf{X}_i, \mathbf{I}\{Y_i = k\})$, $i = 1, \cdots, n$. Corresponding to each partition $\Pi = \{A_1, \cdots, A_{K^{M_n}}\} \in \Omega_n$ define Π^* by

$$\Pi^* = \Pi \times \{0,1\} = \Big\{A_1 \times \{0\}, \cdots, A_{K^{M_n}} \times \{0\}\Big\} \cup \Big\{A_1 \times \{1\}, \cdots, A_{K^{M_n}} \times \{1\}\Big\}.$$

Associated with each Ω_n define the family Ω_n^* of partitions of $\mathbf{B}^d \times \{0,1\}$ by

$$\Omega_n^* = \{\Pi^*\} = \Big\{\Pi \times \{0,1\} \mid \Pi \in \Omega_n\Big\}.$$

Then one has

$$\begin{aligned}
D_n &= \sum_{A\in\Pi_n} \int_A \Big|\bar{P}_k(\widehat{\mathbf{W}}(\mathbf{x})) - \hat{P}_k(\widehat{\mathbf{W}}(\mathbf{x}))\Big| \ \mu(d\mathbf{x}) \\
&\le \sum_{A\in\Pi_n} \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi_n} \frac{\mu(A)}{\mu\Big(A \cup \bigcup_{\ell=1}^N A_{n,m_\ell}\Big)}
\end{aligned}$$

$$\times \Big| n^{-1}\sum_{i=1}^{n} \mathbf{I}\{Y_i = k\}\mathbf{I}\Big\{\mathbf{X}_i \in A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}\Big\}$$
$$- \mathbf{E}\Big(\mathbf{I}\{Y = k\}\mathbf{I}\Big\{\mathbf{X} \in A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}\Big\} \mid \mathbf{T}_n\Big)\Big|$$

$$\leq \sum_{A\in\Pi_n} \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi_n} \Big| n^{-1}\sum_{i=1}^{n} \mathbf{I}\{Y_i = k\}\mathbf{I}\Big\{\mathbf{X}_i \in A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}\Big\}$$
$$- \mathbf{E}\Big(\mathbf{I}\{Y = k\}\mathbf{I}\Big\{\mathbf{X} \in A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}\Big\} \mid \mathbf{T}_n\Big)\Big|$$

$$= \sum_{A\in\Pi_n} \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi_n} \Big|\nu_n\Big((A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}) \times \{1\}\Big) - \nu\Big((A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}) \times \{1\}\Big)\Big|$$

$$\leq \sup_{\Pi\in\Omega_n} \sum_{A\in\Pi} \sup_{A_{n,m_1},\cdots,A_{n,m_N} \in \Pi} \Big|\nu_n\Big((A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}) \times \{1\}\Big) - \nu\Big((A \cup \bigcup_{\ell=1}^{N} A_{n,m_\ell}) \times \{1\}\Big)\Big|$$

$$\leq \sup_{\Pi\in\Omega_n} \sum_{A\in\Pi} \Big\{\Big|\nu_n(A \times \{1\}) - \nu(A \times \{1\})\Big| + \sum_{A\in\Pi} \Big|\nu_n(A \times \{1\}) - \nu(A \times \{1\})\Big|\Big\}$$

$$= \sup_{\Pi\in\Omega_n} \Big\{ \sum_{A\in\Pi} \Big|\nu_n(A \times \{1\}) - \nu(A \times \{1\})\Big| + K^{M_n} \sum_{A\in\Pi} \Big|\nu_n(A \times \{1\}) - \nu(A \times \{1\})\Big|\Big\}$$

$$= (K^{M_n} + 1) \sup_{\Pi\in\Omega_n} \sum_{A\in\Pi} \Big|\nu_n(A \times \{1\}) - \nu(A \times \{1\})\Big|.$$

An application of Lugosi and Nobel's [7] Vapnik-Chervonenkis-type inequality for the partition families of the space $\mathbf{B}^d \times \{0, 1\}$ yields (see Lemma 1 in the cited paper):

$$\mathbf{P}\{D_n > \epsilon\} \leq \mathbf{P}\Big\{ \sup_{\Pi\in\Omega_n} \sum_{A\in\Pi} \Big|\nu_n(A \times \{1\}) - \nu(A \times \{1\})\Big| > \frac{\epsilon}{2K^{M_n}}\Big\}$$
$$\leq 4\Delta(\Omega_n, 2n)2^{2K^{M_n}} . \exp\Big(- \frac{n\epsilon^2}{128K^{2M_n}}\Big).$$

Now condition (b) of the theorem together with the Borel-Cantelli lemma implies that $D_n \to$a.s. 0. As for the term $\mathbf{B}_n$, put

$$S^{A[\mathbf{x}]}_{n,M',K} = A_n[\mathbf{x}] \cup \bigcup_{\ell=1}^{M'} \bigcup_{j=1}^{K-1} A^{(j)}_{n,m_\ell}(\mathbf{x}) =: \bigcup_{i=1}^{1+(K-1)M'} A_i(\mathbf{x}),$$

and observe that

$$B_n \leq \int_{\mathbf{B}^d} \frac{1}{\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)} \Big|\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)h^\epsilon(\mathbf{x}) - \mathbf{E}\Big(\mathbf{I}\Big\{\mathbf{X} \in S^{A[\mathbf{x}]}_{n,M',K}\Big\}h^\epsilon(\mathbf{X}) \mid \mathbf{T}_n\Big)\Big|\mu(d\mathbf{x})$$
$$= \int_{\mathbf{B}^d} \frac{1}{\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)} \Big| \int_{S^{A[\mathbf{x}]}_{n,M',K}} h^\epsilon(\mathbf{x})\mu(d\mathbf{z}) - \int_{S^{A[\mathbf{x}]}_{n,M',K}} h^\epsilon(\mathbf{z})\mu(d\mathbf{z})\Big|\mu(d\mathbf{x})$$

$$
\begin{aligned}
&\leq \int_{\mathbf{B}^d} \frac{1}{\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)} \int_{S^{A[\mathbf{x}]}_{n,M',K}} \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\mathbf{z})\Big| \mu(d\mathbf{z})\mu(d\mathbf{x}) \\
&\leq \int_{\mathbf{B}^d} \frac{1}{\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)} \Big[\sum_{A_i(\mathbf{x}) \in S^{A[\mathbf{x}]}_{n,M',K}:\ \mu(A_i(\mathbf{x}))\neq 0} \mu(A_i^-(\mathbf{x})) \frac{1}{\mu(A_i^-(\mathbf{x}))} \\
&\qquad\qquad \times \int_{A_i^-(\mathbf{x})} \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\mathbf{z})\Big| \mu(d\mathbf{z})\Big] \mu(d\mathbf{x}) \\
&\qquad \text{(where } A_i^- \text{ is the closure of the cell } A_i\text{)} \\
&= \int_{\mathbf{B}^d} \frac{1}{\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)} \sum_{A_i(\mathbf{x}) \in S^{A[\mathbf{x}]}_{n,M',K}:\ \mu(A_i(\mathbf{x}))\neq 0} \mu(A_i^-(\mathbf{x})) \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\xi_i(\mathbf{x}))\Big| \mu(d\mathbf{x}), \\
&\qquad \text{(where } \xi_i(\mathbf{x}) \text{ is a point in } A_i^-(\mathbf{x}) \text{ (the Mean Value theorem))} \\
&\leq \int_{\mathbf{B}^d} \frac{1}{\mu\Big(S^{A[\mathbf{x}]}_{n,M',K}\Big)} \max_{1\leq j\leq M'(K-1)+1} \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\xi_j(\mathbf{x}))\Big| \sum_{A_i(\mathbf{x}) \in S^{A[\mathbf{x}]}_{n,M',K}} \mu(A_i(\mathbf{x}))\mu(d\mathbf{x}), \\
&\qquad \text{(since the distribution of } \mathbf{X} \text{ has a density)} \\
&= \int_{\mathbf{B}^d} \max_{1\leq j\leq M'(K-1)+1} \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\xi_j(\mathbf{x}))\Big| \mu(d\mathbf{x}).
\end{aligned}
$$

Let $A \in \Pi_n$, and let $A^{(i)}_{n,m_\ell}$, $i = 1, \cdots, K-1$, $\ell = 1, \cdots, M'$ be the neighboring cells of A in the sense of (7) and (6); also see Remark A following equation (7). Also, let $S^A_{n,M',K} = A \cup \bigcup_{\ell=1}^{M'} \bigcup_{i=1}^{K-1} A^{(i)}_{n,m_\ell}$. The uniform continuity of h^ϵ (on $\mathbf{B}^d$) implies that given $\epsilon > 0$, there is a $\delta > 0$ such that if $\mathrm{diam}(S^A_{n,M',K}) < \delta$, then $\Big|h^\epsilon(\mathbf{y}_1) - h^\epsilon(\mathbf{y}_2)\Big| < \epsilon$ for all $\mathbf{y}_1, \mathbf{y}_2 \in S^A_{n,M',K}$. Thus

$$
\begin{aligned}
B_n &\leq \sum_{A\in\Pi_n:\ \mathrm{diam}(S^A_{n,M',K})\geq\delta} \int_A \max_{1\leq j\leq M'(K-1)+1} \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\xi_j(\mathbf{x}))\Big| \mu(d\mathbf{x}) \\
&\qquad + \sum_{A\in\Pi_n:\ \mathrm{diam}(S^A_{n,M',K})<\delta} \int_A \max_{1\leq j\leq M'(K-1)+1} \Big|h^\epsilon(\mathbf{x}) - h^\epsilon(\xi_j(\mathbf{x}))\Big| \mu(d\mathbf{x}) \\
&\leq 2 \sum_{A\in\Pi_n:\ \mathrm{diam}(S^A_{n,M',K})\geq\delta} \mu(A) + \sum_{A\in\Pi_n:\ \mathrm{diam}(S^A_{n,M',K})<\delta} \epsilon\ \mu(A) \\
&\qquad \text{(since } h^\epsilon \text{ is bounded)} \\
&\leq 2\mu\Big(\Big\{\mathbf{x} \mid \mathrm{diam}(S^{A[\mathbf{x}]}_{n,M',K}) \geq \delta\Big\}\Big) + \epsilon.
\end{aligned}
$$

Let $A^*_n[\mathbf{x}]$ be the unique cell of Π^*_n that contains the point $\mathbf{x}$. A little effort shows that

$$
S^{A[\mathbf{x}]}_{n,M',K} \subset \bigcup_{j_1=1}^{K} \cdots \bigcup_{j_{M'}=1}^{K} \Big\{\mathbf{y} \mid g_{n,m_\ell}(\mathbf{y}) = [g_{n,m_\ell}(\mathbf{x}) + j_\ell] \pmod K,\ \ell = 1, \cdots, M',
$$

$$\text{and } g_{n,m_\ell}(\mathbf{y}) = g_{n,m_\ell}(\mathbf{x}),\ \ell = M'+1,\cdots,M\Big\}$$

$$= \quad A_n^*[\mathbf{x}].$$

This implies that $B_n \leq 2\mu\Big(\Big\{\mathbf{x} \mid \text{diam}(A_n^*[\mathbf{x}]) \geq \delta\Big\}\Big) + \epsilon$. Therefor, by part (c) of the theorem, $\limsup_{n\to\infty} B_n \leq_{\text{a.s.}} \epsilon$. This proves the theorem since we have shown that

$$\limsup_{n\to\infty} \{E_\epsilon + B_n + C_n + D_n\} \leq_{\text{a.s.}} 3\epsilon,$$

for all $\epsilon > 0$.

References.

1. J. Aitchinson, and C. Aitken, Multivariate binary discrimination by the kernel method, *Biometrika.* **63** (1976). 413-420.

2. L. Breiman, Stacked Regression, *Machine Learning*, **24** (1995). 49-64.

3. L. Devroye, L. Györfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," Springer, New York. 1996.

4. P. Hall, On nonparametric multivariate binary discrimination. *Biometrika. 68* (1981). 287-294.

5. M. Hills, 1967. Allocation rules and their error rates. *J. Royal Statist. Soc. B.* **28** (1967). 1-31.

6. M. LeBlanc, and R. Tibshirani, Combining estimates in regression and classification, *J. Amer. Statist. Assoc.* **91** (1996). 1641-1650.

7. G Lugosi, and A.B Nobel, Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* **24** (1996). 687-706.

8. M. Mojirsheibani, An Almost Surely Optimal Combined Classification Rule. Technical Report No. 331. Laboratory for Research in Statistics & Probability, Carleton University, (1999).

9. M. Mojirsheibani, Combining Classifiers via Discretization, *J. Amer. Statist. Assoc.* **94** (1999). 600-609.

10. M. Mojirsheibani, A Consistent Combined Classification Rule. *Statist. Probab. Lett.* **36** (1997). 43-47

11. J. Schürmann, (1996), "Pattern classification. A unified view of statistical and neural approaches," John Wiley, New York 1996.

12. D. Wolpert, Stacked generalization, *Neural Networks,* **5** (1992). 241-259.