Journal of Multivariate Analysis 124 (2014) 42–56

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A statistical view of clustering performance through the theory of U-processes

Stéphan Clémençon

Institut Telecom, LTCI UMR, Telecom ParisTech/CNRS No. 5141, Telecom ParisTech, 46 rue Barrault, Paris, 75634, France

ARTICLE INFO

Article history:
Received 3 January 2013
Available online 28 October 2013

AMS subject classifications:
62H30
68Q32

Keywords:
Cluster analysis
Pairwise dissimilarity
U-process
Empirical risk minimization
Fast rates
Minimax lower bound
Median clustering

ABSTRACT

Many clustering techniques aim at optimizing empirical criteria that are of the form of a U-statistic of degree two. Given a measure of dissimilarity between pairs of observations, the goal is to minimize the *within cluster* point scatter over a class of partitions of the feature space. It is the purpose of this paper to define a general statistical framework, relying on the theory of U-processes, for studying the performance of such clustering methods. In this setup, under adequate assumptions on the complexity of the subsets forming the partition candidates, the *excess of clustering risk* of the empirical minimizer is proved to be of the order $O_{\mathbb{P}}(1/\sqrt{n})$. A lower bound result shows that the rate obtained is optimal in a minimax sense. Based on recent results related to the tail behavior of degenerate U-processes, it is also shown how to establish tighter, and even faster, rate bounds under additional assumptions. Model selection issues, related to the number of clusters forming the data partition in particular, are also considered. Finally, it is explained how the theoretical results developed here can provide statistical guarantees for empirical clustering aggregation.

1. Introduction

In *cluster analysis*, the objective is to segment a dataset into subgroups, such that data points in the same subgroup are more similar to each other (in a sense that will be specified) than to those in other subgroups. Given the wide range of applications of the clustering paradigm, numerous data segmentation procedures have been introduced in the machine-learning literature (see Chapter 14 in [23] and Chapter 8 in [14] for recent overviews of "off-the-shelf" clustering techniques). Whereas the design of clustering algorithms is still receiving much attention in machine-learning (see [48] and the references therein for instance), the statistical study of their performance, with the notable exception of the celebrated K-means approach, see [22,33,34,7,2] and more recently [9] in the functional data analysis setting, may appear to be not sufficiently well-documented in contrast, as pointed out in [45,13]. Indeed, in the K-means situation, the specific form of the criterion (and of its expectation, the *clustering risk*), as well as that of the cells defining the clusters and forming a partition of the feature space (*Voronoi cells*), permits to use, in a straightforward manner, results of the theory of empirical processes in order to control the performance of empirical clustering risk minimizers. Unfortunately, this *center-based* approach does not carry over into more general situations, where the dissimilarity measure is not a square Hilbertian norm anymore, unless one looses the possibility to interpret the clustering criterion as a function of pairwise dissimilarities between the observations (*cf* K*-medians*).

It is the goal of this paper to establish a general statistical framework for investigating clustering performance. The present analysis is based on the observation that many statistical criteria for measuring clustering accuracy are (symmetric) U-statistics (of degree two), functions of a matrix of dissimilarities between pairs of data points. Such statistics have recently received a good deal of attention in the machine-learning literature, insofar as empirical performance measures of predictive

E-mail address: stephan.clemencon@telecom-paristech.fr.

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.10.001

rules in problems such as *statistical ranking* (when viewed as pairwise classification), see [15], or *learning on graphs* [8], are precisely functionals of this type, generalizing sample mean statistics. By means of uniform deviation results for U-processes, the *Empirical Risk Minimization* paradigm (ERM) can be extended to situations where natural estimates of the risk are U-statistics. In this way, we establish here a rate bound of order $O_{\mathbb{P}}(1/\sqrt{n})$ for the excess of clustering risk of empirical minimizers under adequate complexity assumptions on the cells forming the partition candidates (the bias term is neglected in the present analysis). We prove a lower bound result, claiming that this learning rate cannot be improved in the minimax sense, in absence of further assumptions. A linearization technique, combined with sharper tail results in the case of degenerate U-processes is also used in order to show that tighter rate bounds can be obtained under additional assumptions. In addition, conditions related to the underlying data distribution and to the collection of partitions over which ERM is performed, under which *fast rates of convergence* can be established (*i.e.* rates faster than $O_{\mathbb{P}}(1/\sqrt{n})$), are exhibited. It is also shown how to use the upper bounds proved throughout this analysis in order to deal with the problem of automatic model selection, that of selecting the number of clusters in particular, through *clustering risk structural minimization*. The idea is to add a complexity penalization term to the empirical clustering risk criterion in order to select a clustering rule that achieves a nearly optimal balance between "bias" and "variance", so that the model chosen is not necessarily the most complex and data overfitting may be avoided. Ensemble learning methods have been proved efficient in many area of machine-learning, in unsupervised settings in particular. Here, we formulate a (metric-based) concept of (theoretical) *median clustering rule* over a collection of partitions and state results, based on the same statistical techniques, guaranteeing that it can be attained in an asymptotic fashion, when replacing (pseudo-) distances between partitions by their empirical counterparts. We point out that a very preliminary version of this work has been presented at the conference NIPS 2011.

The paper is structured as follows. In Section 2, the notations are set out, a formal description of cluster analysis, from the "pairwise dissimilarity" perspective, is given and the main theoretical concepts involved in the present analysis are briefly recalled. Section 3 states a general minimax lower bound result for the excess of clustering risk. In Section 4, an upper bound for the performance of empirical minimization of the clustering risk is established in the context of general dissimilarity measures. Section 5 shows how to refine the rate bound previously obtained by means of a recent inequality for degenerate U-processes, while Section 6 deals with automatic selection of the optimal number of clusters. Finally, Section 7 revisits the notion of clustering aggregation by means of the approach developed all along the article. Technical proofs are deferred to the Appendix section.

2. Theoretical background

In this section, after a brief description of the probabilistic framework of the study, the general formulation of the clustering objective, based on the notion of dissimilarity between pairs of observations, is recalled and the connection of the problem of investigating clustering performance with the theory of U-statistics and U-processes is highlighted. Concepts pertaining to this theory and involved in the subsequent analysis are next recalled.

2.1. Probabilistic setup and first notations

Here and throughout, $(X_1, \ldots, X_n)$ denotes a sample of i.i.d. random vectors, valued in a high-dimensional feature space $\mathcal{X}$, typically a subset of the Euclidean space $\mathbb{R}^d$ with $d \gg 1$, with common probability distribution $\mu(dx)$. The indicator function of any event $\mathcal{E}$ will be denoted by $\mathbb{I}\{\mathcal{E}\}$, the usual l_p norm on $\mathbb{R}^d$ by $||x||_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$ when $1 \leq p < \infty$ and by $||x||_\infty = \max_{1\leq i\leq d} |x_i|$ in the case $p = \infty$, with $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. When well-defined, the expectation and the variance of a r.v. Z are denoted by $\mathbb{E}[Z]$ and $\mathrm{Var}(Z)$ respectively. The cardinality of any finite set $\mathcal{A}$ is denoted by $\#\mathcal{A}$. We denote by $x_+ = \max(0, x)$ the positive part of any real number x and, finally, for any partition $\mathcal{P}$ of the space $\mathcal{X}$ we denote by $\Phi_{\mathcal{P}} : \mathcal{X}^2 \to \{0, 1\}$ the binary function that indicates whether two elements of $\mathcal{X}$ belong to the same cell of $\mathcal{P}$ or not: $\Phi_{\mathcal{P}}(x, x') = \sum_{\mathcal{C}\in\mathcal{P}} \mathbb{I}\{(x, x') \in \mathcal{C}^2\}$, for all $(x, x') \in \mathcal{X}^2$.

2.2. Cluster analysis and pairwise dissimilarity

The goal of clustering techniques is to partition the data $(X_1, \ldots, X_n)$ into a given finite number of groups, $K \ll n$ say, so that the observations lying in a same group are more similar to each other than to those in other groups. When equipped with a (Borelian) measure of dissimilarity $D : \mathcal{X}^2 \to \mathbb{R}_+^*$, the clustering task can be rigorously cast as the problem of minimizing the criterion

$$\widehat{W}_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} D(X_i, X_j) \cdot \Phi_{\mathcal{P}}(X_i, X_j), \tag{1}$$

over all possible partitions $\mathcal{P} = \{\mathcal{C}_k : \ 1 \leq k \leq K\}$ of the feature space $\mathcal{X}$. The quantity (1) is generally called the *intra-cluster similarity* or the *within cluster point scatter*. The function D aiming at measuring dissimilarity between pairs of observations, we suppose that it fulfills the following properties:

- (Symmetry) For all $(x, x') \in \mathcal{X}^2$, $D(x, x') = D(x', x)$.
- (Separation) For all $(x, x') \in \mathcal{X}^2$: $D(x, x') = 0, \Leftrightarrow x = x'$.

Typical choices for the dissimilarity measure are of the form $D(x, x') = \phi(||x - x'||_p)$, where $p \geq 1$ and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing function such that $\phi(0) = 0$ and $\phi(t) > 0$ for all $t > 0$. This includes the so-termed "standard K-means" setup, where the dissimilarity measure coincides with the square Euclidean norm (in this case, $p = 2$ and $\phi(t) = t^2$ for $t \geq 0$). Notice that the expectation of the r.v. (1) is equal to the following quantity:

$$W_\mu(\mathcal{P}) = \mathbb{E}\left[D(X, X') \cdot \Phi_{\mathcal{P}}(X, X')\right], \tag{2}$$

where (X, X') denotes a pair of independent r.v.'s drawn from $\mu(dx)$. It will be referred to as the *clustering risk* of the partition $\mathcal{P}$, while its statistical counterpart (1) will be called the *empirical clustering risk*. *Optimal partitions* of the feature space $\mathcal{X}$ are defined as those that minimize $W_\mu(\mathcal{P})$. When no confusion about the distribution $\mu(dx)$ under study is possible, the subscript will be omitted and we will simply write $W(\mathcal{P})$. Before formulating the empirical clustering risk minimization problem, we collect several remarks below.

Remark 1 (*Maximization Formulation*)**.** It is well-known that minimizing the empirical clustering risk (1) is equivalent to maximizing the *between-cluster* scatter point, which is given by $1/(n(n-1)) \cdot \sum_{i,j} D(X_i, X_j) \cdot (1 - \Phi_{\mathcal{P}}(X_i, X_j))$, the sum of these two statistics being independent from the partition $\mathcal{P}$ considered, equal to the quantity $1/(n(n-1)) \cdot \sum_{i \neq j} D(X_i, X_j)$. Its expectation is $\delta_D(\mu) = \int\int D(x, x')\mu(dx)\mu(dx')$, which can be viewed as a *dispersion measure*, extending well-known concepts in the 1-d setting such as the variance or the Gini mean difference.

Remark 2 (*Monotonicity*)**.** Let $\mathcal{P}$ be a partition of $\mathcal{X}$. We call "subpartition of $\mathcal{P}$" any partition $\mathcal{P}'$ of $\mathcal{X}$ whose cells can be obtained by splitting those of $\mathcal{P}$, *i.e.* such that: $\forall \mathcal{C}' \in \mathcal{P}'$, $\exists \mathcal{C} \in \mathcal{P}$ such that $\mathcal{C}' \subset \mathcal{C}$. We then write $\mathcal{P}' \subset \mathcal{P}$. Notice that, in such a case, we necessarily have $\#\mathcal{P} \leq \#\mathcal{P}'$, $\Phi_{\mathcal{P}'}(.,.) \leq \Phi_{\mathcal{P}}(.,.)$ and thus: $W_\mu(\mathcal{P}') \leq W_\mu(\mathcal{P})$ for any probability distribution μ. For completeness, observe however that one may exhibit distributions μ and partitions $\mathcal{P}$, $\mathcal{P}'$ for which we have $W_\mu(\mathcal{P}) < W_\mu(\mathcal{P}')$, whereas $\#\mathcal{P} < \#\mathcal{P}'$. Considering for instance the situation where μ is the uniform distribution on $[0, 1]$ and $D(x, x') = |x - x'|$ for $(x, x') \in [0, 1]^2$, one may easily check that the clustering risk of the partition $\{[0, 1/4[\cup [3/4, 1],\ [1/4, 1/4 + \epsilon[,\ [1/4 + \epsilon, 3/4]\}$ is larger than that of the partition made of two cells only, $[0, 1/2[$ and $[1/2, 1]$.

Remark 3 (*Uniqueness*)**.** By symmetry arguments, one may easily see that, whenever it is attained, the clustering risk minimum over a given collection of partitions is not necessarily achieved in a unique fashion. Considering the uniform distribution on the unit square $[0, 1]^2$, notice for instance that the partitions $\{[0, 1] \times [0, 1/2[, [0, 1] \times [1/2, 1]\}$ and $\{[0, 1/2[\times [0, 1], [1/2, 1] \times [0, 1]\}$ have exactly the same clustering risk, with a function of the square Euclidean norm as dissimilarity measure say.

Suppose we are given a (hopefully sufficiently rich) class Π of partitions of the feature space $\mathcal{X}$. Here, as the distribution μ is unknown in practice, we consider minimizers of the empirical risk $\widehat{W}_n$ over Π, *i.e.* partitions $\widehat{\mathcal{P}}_n^*$ in Π such that

$$\widehat{W}_n\left(\widehat{\mathcal{P}}_n^*\right) = \min_{\mathcal{P} \in \Pi} \widehat{W}_n(\mathcal{P}). \tag{3}$$

The design of practical algorithms for computing (approximately) empirical clustering risk minimizers is beyond the scope of this paper (refer to [23] or to [14] for an overview of "off-the-shelf" clustering methods). Here, focus is on the performance of such empirically defined rules solely.

2.3. U-statistics and U-processes

The subsequent analysis crucially relies on the fact that the quantity (1) that one seeks to optimize is a U-statistic. For clarity's sake, we recall the definition of this class of statistics, generalizing basic sample means.

Definition 1 (*U-Statistic of Degree Two*)**.** Let $X_1, \ldots, X_n$ be independent copies of a random vector X drawn from a probability distribution $\mu(dx)$ on the space $\mathcal{X}$ and $\mathcal{K} : \mathcal{X}^2 \to \mathbb{R}$ be a symmetric function such that $\mathcal{K}(X_1, X_2)$ is square integrable. By definition, the functional

$$U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathcal{K}(X_i, X_j) \tag{4}$$

is a (symmetric) U-statistic of degree two, with kernel $\mathcal{K}$. It is said to be degenerate when $\mathcal{K}^{(1)}(x) \overset{\text{def}}{=} \mathbb{E}[\mathcal{K}(x, X)] = 0$ with probability one for all $x \in \mathcal{X}$, non degenerate otherwise.

The statistic (4) is a natural (unbiased) estimate of the quantity

$$\theta = \int\int \mathcal{K}(x, x')\mu(dx)\mu(dx').$$

The class of U-statistics is very large and include most dispersion measures, including the variance or the Gini mean difference (with $\mathcal{K}(x,x') = (x-x')^2$ and $\mathcal{K}(x,x') = |x-x'|$ respectively, $(x,x') \in \mathbb{R}^2$), as well as the celebrated Wilcoxon location test statistic (with $\mathcal{K}(x,x') = \mathbb{I}\{x+x' > 0\}$ for $(x,x') \in \mathbb{R}^2$ in this case). Although the dependence structure induced by the summation over all pairs of observations makes its study more difficult than that of basic sample means, this estimator has nice properties. It is well-known folklore in mathematical statistics that it is the most *efficient* estimator among all unbiased estimators of the parameter θ (*i.e.* that with minimum variance), see [42]. Precisely, when non degenerate, it is asymptotically normal with limiting variance $4 \cdot \mathrm{Var}(\mathcal{K}^{(1)}(X))$ (refer to Chapter 5 in [36] for an account of asymptotic analysis of U-statistics). As shall be seen in Section 5, the reduced variance property of U-statistics is crucial, when it comes to establish tight rate bounds.

Going back to the U-statistic of degree two (1) estimating (2), observe that its symmetric kernel is: $\forall (x,x') \in \mathcal{X}^2$,

$$\mathcal{K}_{\mathcal{P}}(x,x') = D(x,x') \cdot \Phi_{\mathcal{P}}(x,x') = \sum_{k=1}^{K} D(x,x') \cdot \mathbb{I}\{(x,x') \in \mathcal{C}_k^2\}. \tag{5}$$

Assuming that $\mathbb{E}[D^2(X_1,X_2) \cdot \mathbb{I}\{(X_1,X_2) \in \mathcal{C}_k^2\}] < \infty$ for all $k \in \{1,\ldots,K\}$ and placing ourselves in the situation where $K \geq 1$ is less than $\mathcal{X}$'s cardinality, the U-statistic (1) is always non degenerate, except in the (sole) case where μ's support, $\mathrm{supp}(\mu)$, is made of K elements exactly and all $\mathcal{P}$'s cells are singletons. Indeed, for all $x \in \mathcal{X}$, denoting by $k(x)$ the index of $\{1,\ldots,K\}$ such that $x \in \mathcal{C}_{k(x)}$, we have:

$$\mathcal{K}_{\mathcal{P}}^{(1)}(x) \stackrel{\text{def}}{=} \mathbb{E}[\mathcal{K}_{\mathcal{P}}(x,X)] = \int_{x' \in \mathcal{C}_{k(x)}} D(x,x')\mu(dx'). \tag{6}$$

As the separation property is fulfilled by D, the quantity above is zero iff $C_{k(x)} \cap \mathrm{supp}(\mu) = \{x\}$. In the non degenerate case, notice finally that the asymptotic variance of $\sqrt{n}\{\widehat{W}_n(\mathcal{P}) - W(\mathcal{P})\}$ is equal to $4 \cdot \mathrm{Var}(D(X, \mathcal{C}_{k(X)}))$, where we set $D(x,C) = \int_{x' \in C} D(x,x')\mu(dx')$ for all $x \in \mathcal{X}$ and any measurable set $C \subset \mathcal{X}$.

By definition, a U-process is a collection of U-statistics, one may refer to [19] for an account of the theory of U-processes. Echoing the role played by the theory of empirical processes in the study of the ERM principle in binary classification, the control of the fluctuations of the U-process

$$\left\{\widehat{W}_n(\mathcal{P}) - W(\mathcal{P}) : \ \mathcal{P} \in \Pi\right\}$$

indexed by a set Π of partition candidates will naturally lie at the heart of the present analysis. As shall be seen below, this can be achieved mainly by the means of the Hoeffding representations of U-statistics, see [24].

3. A minimax lower bound

In this section, the goal pursued is to obtain a minimax lower bound for the excess of clustering risk. Let Π be a set of partitions of the feature space $\mathcal{X}$. A data sample $\mathcal{D}_n = \{X_1,\ldots,X_n\}$ made of i.i.d. realizations of the distribution $\mu(dx)$ is here used exclusively to select a partition $\mathcal{P}_n$ from Π, whose clustering performance is measured by the difference $W(\mathcal{P}_n) - W^*$, where $W^* = \inf_{\mathcal{P} \in \Pi} W(\mathcal{P})$. We investigate the problem of finding a lower bound for

$$\sup \mathbb{E}\left[W(\mathcal{P}_n) - W^*\right],$$

where the supremum is taken over all possible distributions $\mu(dx)$ in a nonparametric class $\mathcal{L}$, no matter the selection method considered for choosing $\mathcal{P}_n$ in Π: a worst-case result of this nature says that whatever the method used for picking a partition in Π, there always exists a distribution in $\mathcal{L}$ so that, when data are drawn from the latter, the method performs worse than the bound in expectation. One may refer to Chapter 14 in [18] for lower bounds in the classification context and to Chapter 2 of [40] for a complete description of possible proof techniques leading to such results in a variety of settings. For simplicity's sake, in this section we restrict our attention to the case where $D(x,x') = ||x-x'||_p$ with $1 \leq p \leq +\infty$. By examining the proof given in the Appendix, one may easily see that the lower bound result stated below can be extended to a more general framework.

Theorem 1. *Let Π_K be the class of partitions of the feature space $\mathcal{X} \subset [0,1]^d$ with $K \geq 2$ cells. Let $\mathcal{L}$ be the set of distributions $\mu(dx)$ with support included in $\mathcal{X} \subset [0,1]^d$. Then, for every partition $\mathcal{P}_n \in \Pi_K$ whose choice is based on the sample $X_1,\ldots,X_n$, we have:*

$$\sup_{\mu \in \mathcal{L}} \mathbb{E}_\mu[W_\mu(\mathcal{P}_n) - \inf_{\mathcal{P} \in \Pi_K} W_\mu(\mathcal{P})] \geq C \cdot \sqrt{\frac{1}{n}},$$

where $C > 0$ denotes some constant depending on K and d solely.

The technical proof is inspired from the argument of Theorem 1 in [6] in the K-means context, refer to the Appendix section for further details.

4. A bound for the excess of clustering risk

Here we establish an upper bound for the performance of an empirical minimizer of the clustering risk over a class Π_K of partitions of $\mathcal{X}$ with $K \geq 2$ cells, K being fixed here and supposed to be smaller than $\mathcal{X}$'s cardinality. Of course, the results stated in this section could be extended to situations where the partitions considered do not all count the same number of cells. This restriction is motivated by the fact that many clustering techniques require to fix K in advance and by the use of explicit bounds involving K for automatically selecting the optimal number of cells in Section 6. We denote by W_K^* the clustering risk minimum over all partitions of $\mathcal{X}$ with K cells. The following global suprema of empirical Rademacher averages, characterizing the complexity of the cells forming the partition candidates, shall be involved in the subsequent rate analysis: $\forall n \geq 2$,

$$\mathcal{A}_{K,n} = \sup_{\substack{\mathcal{C}\in\mathcal{P} \\ \mathcal{P}\in\Pi_K}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \epsilon_i D(X_i, X_{i+\lfloor n/2 \rfloor}) \mathbb{I}\{(X_i, X_{i+\lfloor n/2 \rfloor}) \in \mathcal{C}^2\} \right| \tag{7}$$

where $\epsilon = (\epsilon_i)_{i\geq 1}$ is a Rademacher chaos, independent from the X_i's, see [27].

The following theorem reveals that the clustering performance of the empirical minimizer (3) is of the order $O_{\mathbb{P}}(1/\sqrt{n})$, when neglecting the bias term (depending on the richness of Π_K solely). As shown by Theorem 1, this rate bound is actually tight.

Theorem 2. *Consider a class Π_K of partitions with $K \geq 1$ cells and suppose that:*

- *there exists $B < \infty$ such that for all $\mathcal{P}$ in Π_K, any $\mathcal{C}$ in $\mathcal{P}$,*

$$\sup_{(x,x')\in\mathcal{C}^2} D(x, x') \leq B,$$

- *the expectation of the Rademacher average $\mathcal{A}_{K,n}$ is of the order $O(n^{-1/2})$.*

Let $\delta > 0$. For any empirical clustering risk minimizer $\widehat{\mathcal{P}}_n^*$, we have with probability at least $1-\delta$:

$$\begin{aligned} \forall n \geq 2, \quad W(\widehat{\mathcal{P}}_n^*) - W_K^* &\leq 4K\mathbb{E}[\mathcal{A}_{K,n}] + 2BK\sqrt{\frac{2\log(1/\delta)}{n}} + \left(\inf_{\mathcal{P}\in\Pi_K} W(\mathcal{P}) - W_K^*\right) \\ &\leq c(B,\delta)\cdot\frac{K}{\sqrt{n}} + \left(\inf_{\mathcal{P}\in\Pi_K} W(\mathcal{P}) - W_K^*\right), \end{aligned} \tag{8}$$

for some constant $c(B,\delta) < \infty$, independent from n and K.

The key for proving (8) is to express the U-statistic $\widehat{W}_n(\mathcal{P})$ in terms of sums of i.i.d. r.v.'s, as that involved in the Rademacher average (7):

$$\widehat{W}_n(\mathcal{P}) = \frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathcal{K}_{\mathcal{P}}(X_i, X_{i+\lfloor n/2 \rfloor}), \tag{9}$$

where the average is taken over $\mathfrak{S}_n$, the symmetric group of order n. The main point lies in the fact that standard techniques in empirical process theory can be then used to control $\widehat{W}_n(\mathcal{P}) - W(\mathcal{P})$ uniformly over Π_K under adequate hypotheses, see the proof in the Appendix for technical details. We underline that, naturally, the complexity assumption is also a crucial ingredient of the result stated above, and more generally to clustering consistency results, see Example 1 in [13]. We also point out that the ERM approach is by no means the sole method to obtain error bounds in the clustering context. Just like in binary classification (see [28]), one may use a notion of *stability* of a clustering algorithm to establish such results, see [44,37] and the references therein. Refer to [46,47] for error bounds proved through the stability approach in the standard K-means setup. In addition, we emphasize that the simplifying boundedness hypothesis $\sup\{D(x,x') : (x,x') \in \mathcal{C}^2, \mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi\} < +\infty$ can be easily relaxed. Indeed, under adequate tail assumptions for $D(X, X')$ for instance, rate bounds can be obtained by adapting slightly the argument given in the Appendix section by means of the standard *truncation trick*, originally introduced in [21].

Before showing how the bound for the excess of risk stated above can be improved under stronger assumptions, a few remarks are in order.

Remark 4 (*On the Complexity Assumption*)**.** We point out that standard entropy metric arguments can be used in order to bound the expected value of the Rademacher average $\mathcal{A}_n$, see [10] for instance. In particular, if the set of functions $\mathcal{F}_{\Pi_K} = \{(x,x') \in \mathcal{X}^2 \mapsto D(x,x')\cdot\mathbb{I}\{(x,x')\in\mathcal{C}^2\} : \mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi_K\}$ is a VC major class with finite VC dimension V (see [20]), then $\mathbb{E}[\mathcal{A}_{K,n}] \leq c\sqrt{V/n}$ for some universal constant $c < \infty$. This covers a wide variety of situations, including the case where $D(x,x') = ||x-x'||_p^\beta$ and the class of sets $\{\mathcal{C} : \mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi_K\}$ is of finite VC dimension.

Remark 5 (*K-Means*)**.** In the standard K-means approach, the dissimilarity measure is $D(x, x') = ||x - x'||_2^2$ and partition candidates are indexed by a collection c of distinct "centers" $c_1, \ldots, c_K$ in $\mathcal{X}$: $\mathcal{P}_c = \{C_1, \ldots, C_K\}$ with $C_k = \{x \in \mathcal{X} : ||x - c_k||_2 = \min_{1 \leq l \leq K} ||x - c_l||_2\}$ for $1 \leq k \leq K$ (with adequate distance-tie breaking). One may easily check that for this specific collection of partitions Π_K and this choice for the dissimilarity measure, the class $\mathcal{F}_{\Pi_K}$ is a VC major class with finite VC dimension, see Section 19.1 in [18] for instance. Additionally, it should be noticed than in most practical clustering procedures, center candidates are picked in a data-driven fashion, being taken as the averages of the observations lying in each cluster/cell. In this respect, the M-estimation problem formulated here can be considered to a certain extent as closer to what is actually achieved by K-means clustering techniques in practice, than the usual formulation of the K-means problem (as an optimization problem over $c = (c_1, \ldots, c_K)$ namely).

Remark 6 (*Weighted Clustering Criteria*)**.** Notice that, in practice, the measure D involved in (1) may depend on the data. For scaling purpose, one could assign data-dependent weights $\omega = (\omega_i)_{1 \leq i \leq d}$ in a coordinatewise manner, leading to $\widehat{D}(x, x') = \sum_{i=1}^d (x_i - x_i')^2 / \widehat{\sigma}_i^2$ for instance, where $\widehat{\sigma}_i^2$ denotes the sample variance related to the ith coordinate. Although the criterion reflecting the performance is not a U-statistic anymore, the theory we develop here can be straightforwardly used for investigating clustering accuracy in such a case. Indeed, it is easy to control the difference between the latter and the U-statistic (1) with $D(x, x') = \sum_{i=1}^d (x_i - x_i')^2 / \sigma_i^2$, the σ_i^2's denoting the theoretical variances of μ's marginals, under adequate moment assumptions.

In the analysis carried out here we do not investigate the order of magnitude of the bias term. Nevertheless, we point out that, in some situations, it can be however quantified under adequate hypotheses. For instance, consider the case where $\mathcal{X} = [0, 1]^d$ and the cells of Π_K's elements can all be obtained by binding together hypercubes of side length 2^{-j}, with $j \geq 1$. We denote by $\mathcal{H}_j$ the set of all such hypercubes (it is of cardinality 2^{jd}) and set $\Pi_K = \Pi_{j,K}$ in this case. Assume in addition that $\mu(dx)$ has a bounded density with respect to Lebesgue measure and there exists an optimal partition $\mathcal{P}^* = \{\mathcal{C}_k^* : 1 \leq k \leq K\}$, whose cells have boundaries that are all of finite perimeter, *i.e.* $per(\partial \mathcal{C}_k^*) < \infty$ for all $k \in \{1, \ldots, K\}$. Then, one may easily show that

$$\inf_{\mathcal{P} \in \Pi_{j,K}} W(\mathcal{P}) - W^* \leq c ||d\mu/dx||_\infty \cdot \max_{H \in \mathcal{H}_j} \sup_{(x,x') \in H^2} D(x, x') \cdot \max_{1 \leq k \leq K} per(\partial \mathcal{C}^*) \cdot 2^{-jd},$$

for some constant $c < +\infty$, see Proposition 9.7 in [29]. Hence, in the case where $D(x, x') = ||x - x'||_\infty^\gamma$ with $\gamma > 0$ for instance, the bias term is of order $2^{-j(\gamma+d)}$, while the stochastic term can be easily shown to be of order $2^{jd}/\sqrt{n}$ (using classically a simple union bound argument, the number of possible cells being finite, less than $2^{2^{jd}}$, in this case). Choosing the level of resolution $j = j(n)$ so that $2^{j(n)} \sim n^{1/(4d+2\gamma)}$ as $n \to \infty$ yields a rate bound of order $n^{-(\gamma+d)/(2\gamma+4d)}$ in (8).

5. Tighter bounds for empirical clustering risk minimizers

We now show that one may refine the rate bound established above, by considering another representation of the U-statistic (1), its (second) *Hoeffding decomposition*, see [36].

5.1. Improved first order analysis

The main argument of the subsequent analysis lies in the following orthogonal decomposition: for all partition $\mathcal{P}$,

$$\widehat{W}_n(\mathcal{P}) - W(\mathcal{P}) = 2L_n(\mathcal{P}) + M_n(\mathcal{P}), \tag{10}$$

$L_n(\mathcal{P}) = (1/n) \sum_{i=1}^n \sum_{\mathcal{C} \in \mathcal{P}} \mathcal{H}_{\mathcal{C}}^{(1)}(X_i)$ being a simple average of i.i.d. random variables with, for $(x, x') \in \mathcal{X}^2$,

$$\mathcal{H}_{\mathcal{C}}(x, x') = D(x, x') \cdot \mathbb{I}\{(x, x') \in \mathcal{C}^2\} \quad \text{and} \quad \mathcal{H}_{\mathcal{C}}^{(1)}(x) = D(x, \mathcal{C}) \cdot \mathbb{I}\{x \in \mathcal{C}\} - D(\mathcal{C}, \mathcal{C}),$$

where $D(\mathcal{C}, \mathcal{C}) = \int_{x \in \mathcal{C}} D(x, \mathcal{C}) \mu(dx)$ and $\mathbb{E}[\mathcal{H}_{\mathcal{C}}(x, X)] = D(x, \mathcal{C}) \cdot \mathbb{I}\{x \in \mathcal{C}\}$, and $M_n(\mathcal{P})$ being a degenerate U-statistic based on the X_i's with kernel given by: $\sum_{\mathcal{C} \in \mathcal{P}} \mathcal{H}_{\mathcal{C}}^{(2)}(x, x')$, where

$$\mathcal{H}_{\mathcal{C}}^{(2)}(x, x') = \mathcal{H}_{\mathcal{C}}(x, x') - \mathcal{H}_{\mathcal{C}}^{(1)}(x) - \mathcal{H}_{\mathcal{C}}^{(1)}(x') - D(\mathcal{C}, \mathcal{C}),$$

for all $(x, x') \in \mathcal{X}^2$. The leading term in (10) is the (centered) sample mean $2L_n(\mathcal{P})$, of the order $O_{\mathbb{P}}(\sqrt{1/n})$, while the second term is of the order $O_{\mathbb{P}}(1/n)$. Hence, provided this holds true uniformly over $\mathcal{P}$, the main contribution to the rate bound should arise from the quantity

$$\sup_{\mathcal{P} \in \Pi_K} |2L_n(\mathcal{P})| \leq 2K \sup_{\mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi_K} \left| (1/n) \sum_{i=1}^n \mathcal{H}_{\mathcal{C}}^{(1)}(X_i) - D(\mathcal{C}, \mathcal{C}) \right|,$$

which thus leads to consider the following suprema of empirical Rademacher averages:

$$\mathcal{R}_{K,n} = \sup_{\mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi_K} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i D(X_i, \mathcal{C}) \cdot \mathbb{I}\{X_i \in \mathcal{C}\} \right|. \tag{11}$$

This supremum clearly has smaller mean and variance than (7). We also introduce the quantities:

$$Z_\epsilon = \sup_{\mathcal{C}\in\mathcal{P},\mathcal{P}\in\Pi_K} \left| \sum_{i,j} \epsilon_i \epsilon_j \mathcal{H}_{\mathcal{C}}^{(2)}(X_i, X_j) \right|,$$

$$U_\epsilon = \sup_{\mathcal{C}\in\mathcal{P},\mathcal{P}\in\Pi_K} \sup_{\alpha:\,\sum_j \alpha_j^2 \le 1} \sum_{i,j} \epsilon_i \alpha_j \mathcal{H}_{\mathcal{C}}^{(2)}(X_i, X_j),$$

$$M = \sup_{\mathcal{C}\in\mathcal{P},\mathcal{P}\in\Pi_K} \sup_{1\le j\le n} \left| \sum_{i} \epsilon_i \mathcal{H}_{\mathcal{C}}^{(2)}(X_i, X_j) \right|.$$

Theorem 3. *Consider a class Π_K of partitions with $K \geq 1$ cells and suppose that:*

- *there exists $B < \infty$ such that, for all $\mathcal{P} \in \Pi_K$, $\mathcal{C} \in \mathcal{P}$,*

$$\sup_{(x,x')\in\mathcal{C}^2} D(x, x') \leq B.$$

Let $\delta > 0$. For any empirical clustering risk minimizer $\widehat{\mathcal{P}}_n^*$, with probability at least $1 - \delta$: $\forall n \geq 2$,

$$W(\widehat{\mathcal{P}}_n^*) - W_K^* \leq 4K\mathbb{E}[\mathcal{R}_{K,n}] + 2BK\sqrt{\frac{\log(2/\delta)}{n}} + K\kappa(n,\delta) + \left(\inf_{\mathcal{P}\in\Pi_K} W(\mathcal{P}) - W_K^*\right), \tag{12}$$

where we set for some universal constant $C < \infty$, independent from n, N and K:

$$\kappa(n,\delta) = C\left(\mathbb{E}[Z_\epsilon] + \sqrt{\log(1/\delta)}\mathbb{E}[U_\epsilon] + (n + \mathbb{E}[M])/\log(1/\delta)\right)/n^2. \tag{13}$$

The result above relies on the moment inequality for degenerate U-processes proved in [15].

Remark 7 (*Localization*)**.** The same argument can be used to decompose $\Lambda_n(\mathcal{P}) - \Lambda(\mathcal{P})$, where $\Lambda_n(\mathcal{P}) = \widehat{W}_n(\mathcal{P}) - W_K^*$ is an estimate of the excess of risk $\Lambda(\mathcal{P}) = W(\mathcal{P}) - W_K^*$, and, by means of concentration inequalities, to obtain next a sharp upper bound that involves the modulus of continuity of the variance of the Rademacher average indexed by the convex hull of the set of functions $\{\sum_{\mathcal{C}\in\mathcal{P}} D(x,\mathcal{C})\cdot\mathbb{I}\{x\in\mathcal{C}\} - \sum_{\mathcal{C}^*\in\mathcal{P}^*} D(x,\mathcal{C}^*)\cdot\{x\in\mathcal{C}^*\} :\ \mathcal{P}\in\Pi_K\}$, following in the footsteps or recent advances in binary classification, see [27] and subsection 5.3 in [10].

5.2. Fast rates of convergence

We now show that one may tighten the rate bound established above in specific (but not uncommon) situations, just like in binary classification [31,41,27] or in ranking [15–17] under so-termed *low-noise* conditions. In the unsupervised setup, the conditions we require to establish faster rates of convergence are of completely different nature. Let Π be a collection of partitions of the space $\mathcal{X}$.

Conditions FR(α). Suppose that the following assumptions are fulfilled.

(i) There exists $\mathcal{P}^*$ in Π such that $W(\mathcal{P}^*) = W^*$.
(ii) There exists $B < \infty$ such that, for all $\mathcal{P} \in \Pi$, $\mathcal{C} \in \mathcal{P}$,

$$\sup_{(x,x')\in\mathcal{C}^2} D(x, x') \leq B.$$

(iii) There exist $\alpha \in [0, 1]$ and $\kappa < \infty$ such that: $\forall \mathcal{P} \in \Pi$, $\forall x \in \mathcal{X}$,

$$\mathbb{E}\left[\mathbb{I}\{\Phi_{\mathcal{P}}(x, X)\} \neq \{\Phi_{\mathcal{P}^*}(x, X)\}\right] \leq \kappa \cdot \left(W(\mathcal{P}) - W^*\right)^\alpha. \tag{14}$$

Observe that, when $\alpha = 0$, condition (iii) above is void. In contrast, when $\alpha > 0$, it ensures that, as $W(\mathcal{P})$ gets closer to W^*, $\mathcal{P}$ gets closer to $\mathcal{P}^*$. In particular, this condition guarantees uniqueness of the minimizer in Π and, truth should be said, maybe hard to check in practice. Notice also that the restriction $\alpha \leq 1$ arises from the fact that, for any $\mathcal{P} \in \Pi$,

$$W(\mathcal{P}) - W^* \leq B \cdot \mathbb{E}\left[\mathbb{I}\{\Phi_{\mathcal{P}}(X, X')\} \neq \{\Phi_{\mathcal{P}^*}(X, X')\}\right].$$

The next result reveals that, under the set of conditions **FR**(α) with $\alpha \in [0, 1]$, the rate attained by the empirical minimizers of the clustering risk is of the order $O_{\mathbb{P}}(n^{-1/(2-\alpha)})$ (when neglecting the bias term), which clearly improves upon the rate stated in Theorem 2 as soon as $\alpha > 0$. As in [15], the argument crucially relies on the small variance property of the U-statistics $\widehat{W}_n(\mathcal{P})$, $\mathcal{P} \in \Pi$, that empirically reflect clustering performance.

Theorem 4. *Consider a class Π of partitions. Suppose that it is of cardinality $N < +\infty$ and that conditions* **FR**(α) *are satisfied. Let $\delta > 0$. For any empirical clustering risk minimizer $\widehat{\mathcal{P}}_n^*$, we have with probability at least $1-\delta$:*

$$\forall n \geq 2, \quad W(\widehat{\mathcal{P}}_n^*) - W^* \leq c \cdot \left(\frac{\log(N/\delta)}{n}\right)^{1/(2-\alpha)}, \tag{15}$$

for some constant $c < \infty$, independent from n and N.

Remark 8 (*On the Complexity Assumption (Bis)*)**.** Our major concern is here to show how the conditions **FR**(α) makes the clustering problem easier from a statistical perspective, rather than to state rate bound results in the full generality. For this reason and to avoid a lengthy technical analysis, a restrictive setup, stipulating the finiteness of the collection of partition candidates, is considered. We point tout that the result above extends to a much more general framework, including the case where complexity assumptions are expressed in terms of finite VC dimension or through (conditional) Rademacher averages, just like in Theorem 5 of [15] (see Theorem 4's proof in the Appendix for further details). In a similar manner, condition (i) in **FR**(α) could be also classically relaxed at the expense of complications in the proof.

Consider $\mathcal{P} \in \Pi$. The argument for proving Theorem 4 relies on the study of the behavior of the U-statistic $\Lambda_n(\mathcal{P})$ based on the sample $X_1, \ldots, X_n$, with (symmetric) kernel

$$\mathcal{H}_{\mathcal{P}}(x, x') = D(x, x') \cdot \left\{\Phi_{\mathcal{P}}(x, x') - \Phi_{\mathcal{P}^*}(x, x')\right\}.$$

The expected value of the statistic $\Lambda_n(\mathcal{P})$ is equal to the excess of risk $\Lambda(\mathcal{P}) = W(\mathcal{P}) - W^*$ and we have the so-termed *second Hoeffding's representation* (see Chapter 5 in [36]):

$$\Lambda_n(\mathcal{P}) - \Lambda(\mathcal{P}) = 2\mathcal{L}_n(\mathcal{P}) + \mathcal{M}_n(\mathcal{P}), \tag{16}$$

where $\mathcal{L}_n(\mathcal{P}) = (1/n) \cdot \sum_{i \leq n} \mathcal{H}^{(1)}(X_i)$ is a standard average of centered i.i.d. r.v.'s, with $\mathcal{H}^{(1)}(x) = \mathbb{E}[\mathcal{H}_{\mathcal{P}}(X, x)] - W(\mathcal{P})$ for all $x \in \mathcal{X}$, and $\mathcal{M}_n(\mathcal{P})$ is the degenerate U-statistic based on the X_i's with kernel given by: $\forall (x, x') \in \mathcal{X}^2$, $\mathcal{H}^{(2)}(x, x') = \mathcal{H}_{\mathcal{P}}(x, x') - \mathcal{H}^{(1)}(x) - \mathcal{H}^{(1)}(x') - \Lambda(\mathcal{P})$.

The following result shows that $\mathcal{M}_n(\mathcal{P})$ is of order $O_{\mathbb{P}}(1/n)$ uniformly over Π. Incidentally, we point that it may be extended to cases where Π is of infinite cardinality, by means of the Arcones–Giné inequality for degenerate U-processes indexed by a collection of kernels with finite VC dimension or by using the inequality stated in Theorem 11 in [15] in an even more general context, see Lemma 8 in the Appendix.

Lemma 1. *Suppose that Theorem 4's assumptions are fulfilled. There exists a universal constant $c < \infty$ such that for all $\delta \in (0, 1)$, we have with probability at least $1-\delta$: $\forall n \geq 2$,*

$$\max_{\mathcal{P} \in \Pi} |\mathcal{M}_n(\mathcal{P})| \leq c \frac{\log(N/\delta)}{n}.$$

Proof. This is a straightforward application of the Bernstein type exponential inequality for degenerate U-statistics with bounded kernel proved in [3] (see assertion (d) of Proposition 2.3 therein) combined with the union bound. □

The bound stated in the following result provides the other key ingredient. Notice that, for $\alpha = 1$, it echoes the fast rate condition in the K-means context introduced in [2] (see Eq. (8) therein).

Lemma 2 (*Control of the Conditional Variance*)**.** *Under Theorem 4's assumptions, we have for all $\mathcal{P} \in \Pi$,*

$$\mathrm{Var}\left(\mathcal{H}^{(1)}(X)\right) \leq c \left(\Lambda(\mathcal{P})\right)^{\alpha},$$

for some finite constant $c > 0$.

Proof. Taking $c = \kappa B^2$, this straightforwardly results from conditions (ii) and (iii) in **FR**(α). □

We point out finally that the following (consistent) estimate of the clustering risk $W(\mathcal{P})$, obtained by splitting the dataset into two halves, could have been used:

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} D(X_i, X_{i+\lfloor n/2 \rfloor}) \cdot \Phi_{\mathcal{P}}(X_i, X_{i+\lfloor n/2 \rfloor}).$$

However, one would have lost the reduced variance property and deriving fast rates for the empirical minimizer would have required a condition much stronger than condition (iii) in **FR**(α). Namely, it would have led to assume that: $\forall \mathcal{P} \in \Pi$, $\forall (x, x') \in \mathcal{X}^2$,

$$\mathbb{I}\{\Phi_{\mathcal{P}}(x, x') \neq \Phi_{\mathcal{P}^*}(x, x')\} \leq \kappa \cdot \left(W(\mathcal{P}) - W^*\right)^{\alpha}.$$

6. Model selection — choosing the number of clusters

In the previous sections, we investigated the principle of empirical clustering risk minimization, that is selecting a partition from a given class/model Π by minimizing the estimate $\widehat{W}_n(\mathcal{P})$ over all $\mathcal{P} \in \Pi$. This provides a partition whose risk is close to the minimum if Π is (i) rich enough to contain a partition with a low clustering risk and (ii) sufficiently small so that finding the best partition candidate in Π can be statistically guaranteed (see Theorem 2). These two requirements are clearly in conflict and call a data-based procedure which achieves a trade-off and permits to select a good model (not necessarily the most complex) from one of classes $\Pi_1, \Pi_2, \ldots$. We now turn to the problem of *model selection*, see [30]. A crucial issue in data segmentation is to determine the number K of cells that exhibits the most the clustering phenomenon in the data. A variety of automatic procedures for choosing a good value for K have been proposed in the literature, based on data splitting, resampling or sampling techniques [32,39,38]. Here we consider a complexity regularization method that avoids to have recourse to such techniques and uses a data-dependent penalty term based on the analysis carried out above.

Suppose that we have a sequence $\Pi_1, \Pi_2, \ldots$ of collections of partitions of the feature space $\mathcal{X}$ such that, for all $K \geq 1$, the elements of Π_K are made of K cells and fulfill the assumptions of Theorem 2. In order to avoid overfitting, consider the (data-driven) complexity penalty given by

$$\text{PEN}(n, K) = 3K\mathbb{E}_\epsilon[\mathcal{A}_{K,n}] + \frac{27BK\log K}{n} + \sqrt{(2B\log K)/n} \tag{17}$$

and the minimizer $\widehat{\mathcal{P}}_{\widehat{K},n}$ of the penalized empirical clustering risk, with

$$\widehat{K} = \arg\min_{K\geq 1} \left\{\widehat{W}_n(\widehat{\mathcal{P}}_{K,n}) + pen(n, K)\right\} \quad \text{and} \quad \widehat{W}_n(\widehat{\mathcal{P}}_{K,n}) = \min_{\mathcal{P}\in\Pi_K} \widehat{W}_n(\mathcal{P}).$$

The next result shows that the partition thus selected nearly achieves the performance that would be obtained with the help of an oracle, revealing the value of the index K that minimizes $\mathbb{E}[\widehat{\mathcal{P}}_{K,n}] - W^*$, with $W^* = \inf_{\mathcal{P}} W(\mathcal{P})$.

Theorem 5 (*An Oracle Inequality*)**.** *Suppose that, for all $K \geq 1$, the assumptions of Theorem 2 are fulfilled. Then, we have:*

$$\mathbb{E}\left[W(\widehat{\mathcal{P}}_{\widehat{K},n})\right] - W^* \leq \min_{K\geq 1}\left\{W_K^* - W^* + pen(n, K)\right\} + \frac{\pi^2}{6}\left(2B\sqrt{\frac{2}{n}} + \frac{18B}{n}\right). \tag{18}$$

Of course, the penalty could be slightly refined using the results of Section 5. Straightforward extensions are left to the reader.

7. Aggregation of clusterings — the Rand median

The principle of aggregation of simple decision rules have recently lead to very efficient methods, generally referred to as *ensemble learning* techniques. In the context of supervised learning, this approach has been shown to enhance prediction accuracy and stability both at the same time, see [11,12] in binary classification for instance. Several attempts have been made to extend the aggregation paradigm to the unsupervised setting (refer to [1] for instance), but, whereas the design of committee-based classification/regression rules simply relies in general on the computation of a (possibly weighted) average, devising aggregation of clustering rules is in contrast much less straightforward. A possible angle for defining a consensus among clusterings is the so-termed *metric approach*, leading to the notion of *median clustering*, see [26]. Ordinal approaches, involving tournaments, could also been considered for this purpose but are beyond the scope of the present analysis, refer to [5,4] for instance. In this section, we focus on a metric of reference, related to the celebrated *Rand index*, and show how the theory of U-processes again can be used to provide statistical guarantees for *empirical median computation*. We point out that the Rand distance is by no means the sole way of measuring closeness between partitions/clusterings and the results established subsequently could be easily extended to any metric involving pairwise comparisons, see [43].

Consider two partitions $\mathcal{P}$ and $\mathcal{P}'$ of the input space $\mathcal{X}$. A possible way of defining a distance between the latter is to compute the probability that a pair of instances independently drawn from $\mu(dx)$ both belong to a same cell for one partition but not for the other. This leads to the *Rand distance*, given by:

$$d_R(\mathcal{P}, \mathcal{P}') = \mathbb{P}\left\{\Phi_{\mathcal{P}}(X, X') \neq \Phi_{\mathcal{P}'}(X, X')\right\}. \tag{19}$$

In general, the rate of "concording pairs", $1 - d_R(\mathcal{P}, \mathcal{P}')$ namely, is usually referred to as the *Rand index*, see [35]. The statistical counterpart of this quantity based on the sample $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ will be called the *empirical Rand distance*, it is the U-statistic of degree two given by

$$\widehat{d}_R(\mathcal{P}, \mathcal{P}') = \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \mathbb{I}\left\{\Phi_{\mathcal{P}}(X_i, X_j) \neq \Phi_{\mathcal{P}'}(X_i, X_j)\right\}, \tag{20}$$

whose kernel is defined by: $\mathcal{K}_{\mathcal{P},\mathcal{P}'}(x,x') = \mathbb{I}\{\Phi_{\mathcal{P}}(x,x') \neq \Phi_{\mathcal{P}'}(x,x')\}$ for all $(x,x') \in \mathcal{X}^2$. Notice that $\widehat{d}_R(\mathcal{P},\mathcal{P}')$ is an unbiased estimate of $d_R(\mathcal{P},\mathcal{P}')$, asymptotically Gaussian with asymptotic variance given by

$$4 \times \left\{ \sum_{(\mathcal{C},\mathcal{C}')\in\mathcal{P}\times\mathcal{P}'} \mu^2(\mathcal{C}\Delta\mathcal{C}')\mu(\mathcal{C}\cap\mathcal{C}') - \left(\sum_{(\mathcal{C},\mathcal{C}')\in\mathcal{P}\times\mathcal{P}'} \mu(\mathcal{C}\Delta\mathcal{C}')\mu(\mathcal{C}\cap\mathcal{C}') \right)^2 \right\},$$

when $\mathcal{P} \neq \mathcal{P}'$ (see further details in the Appendix). The next result shows that, under adequate moment conditions, two partitions that are close in the sense of the Rand distance have comparable intra-cell similarity. It is a simple consequence of Hölder inequality, the proof is omitted.

Lemma 3 (*On Rand Distance and Intra-Cell Similarity*)**.** *Let $1 < q \leq \infty$. Suppose that:*

$$\left(\mathbb{E}\left[D^q(X,X')\right]\right)^{1/q} < +\infty,$$

where (X,X') denotes a pair of independent r.v.'s drawn from $\mu(dx)$. Then, for any partitions $\mathcal{P}$ and $\mathcal{P}'$ of the feature space $\mathcal{X}$, we have:

$$\left|W(\mathcal{P}) - W(\mathcal{P}')\right| \leq \left(\mathbb{E}\left[D^q(X,X')\right]\right)^{1/q} \cdot \left(d_R(\mathcal{P},\mathcal{P}')\right)^{1-1/q}. \tag{21}$$

Notice that, in Section 5.2, a similar control has already been used and the fast rate condition states that the reverse control holds true when one of the partition is the clustering risk minimizer.

Consensus. Given a finite collection of partitions of the feature space, we now recall how to define a natural notion of *median/central partition* based on the distance introduced above, see [4].

Definition 2 (*Rand Median*)**.** Let $1 \leq M < \infty$ and $\mathbf{P} = \{\mathcal{P}^{(1)},\ldots,\mathcal{P}^{(M)}\}$ be a collection of M partitions of the space $\mathcal{X}$. A *Rand consensus* for $\mathbf{P}$ with respect to a set Π of partitions of $\mathcal{X}$ is any partition $\overline{\mathcal{P}} \in \Pi$ such that:

$$\Theta(\overline{\mathcal{P}},\mathbf{P}) = \inf_{\mathcal{P}\in\Pi} \Theta(\mathcal{P},\mathbf{P}), \tag{22}$$

where $\Theta(\mathcal{P},\mathbf{P}) = \sum_{m=1}^M d_R(\mathcal{P},\mathcal{P}^{(m)})$ for any $\mathcal{P}$ in Π.

Remark 9 (*On Existence and Uniqueness of The Median*)**.** We highlight the fact that, in the general case, there is no guarantee that a consensus partition does exist, *i.e.* that the infimum in (22) is attained. Notice however that, when the set Π is of finite cardinality, medians always exist. Observe also that, when a median exists, it is in general not unique.

Remark 10 (*Alternative Notions of Median Partitions*)**.** We point out that many other ways of quantifying the dissimilarity between clusterings have been proposed in the literature, see [25] or [43]. Focus is here on the Rand consensus, mainly because it is defined in a pairwise manner and can be investigated by means of U-statistic/process tools.

In practice, the distribution μ is unknown and one relies on empirical estimates of the Rand distances. This leads to the notion of *empirical consensus*: for any finite collection of partitions $\mathbf{P} = \{\mathcal{P}^{(m)} : 1 \leq m \leq M\}$ and any set of distributions, an empirical median partition is any partition $\widehat{\overline{\mathcal{P}}}_n \in \Pi$ such that

$$\widehat{\Theta}_n(\widehat{\overline{\mathcal{P}}}_n,\mathbf{P}) = \inf_{\mathcal{P}\in\Pi} \widehat{\Theta}_n(\mathcal{P},\mathbf{P}), \tag{23}$$

where $\widehat{\Theta}_n(\mathcal{P},\mathbf{P}) = \sum_{m=1}^M \widehat{d}_R(\mathcal{P},\mathcal{P}^{(m)})$ for any $\mathcal{P}$ in Π. In contrast to theoretical consensus partitions, empirical medians always exist, insofar as, over the set Π, the function $\widehat{\Theta}_n(.,\mathbf{P})$ takes a finite number of values only. However, its computation is a NP-hard problem and generally requires the use of meta-heuristics, see the account in [26] and the references therein for instance. Here we do not address computational issues and focus on the properties of (empirical) median partitions solely. The following result reveals that empirical median partitions are asymptotically median in the sense of definition (22), under adequate control of the complexity of the set of partitions over which the median is taken.

Proposition 1 (*Empirical Aggregation of Partitions*)**.** *Let $M \geq 1$, $\mathbf{P} = \{\mathcal{P}^{(m)} : 1 \leq m \leq M\}$ and Π be two sets of partitions of $\mathcal{X}$ such that the collection of sets $\{\mathcal{C} : \mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi\}$ is of finite* VC *dimension. For any $\mathcal{P} \in \Pi$, consider*

$$\widehat{\Theta}_n(\mathcal{P},\mathbf{P}) = \sum_{m=1}^M \widehat{d}_R(\mathcal{P},\mathcal{P}^{(m)}),$$

where the empirical Rand distance $\widehat{d}_R(.,.)$ is computed from $n \geq 1$ independent observations drawn from $\mu(dx)$. Assume that $\widehat{\overline{\mathcal{P}}}_n \in \Pi$ is such that $\widehat{\Theta}_n(\widehat{\overline{\mathcal{P}}}_n,\mathbf{P}) = \inf_{\mathcal{P}\in\Pi} \widehat{\Theta}_n(\mathcal{P},\mathbf{P})$. Then, as $n \to \infty$, we have:

$$\Theta(\widehat{\overline{\mathcal{P}}}_n,\mathbf{P}) \to \inf_{\mathcal{P}\in\Pi} \Theta(\mathcal{P},\mathbf{P}) \text{ with probability one}. \tag{24}$$

Additionally, this convergence takes place at the rate $O_{\mathbb{P}}(1/\sqrt{n})$.

The proof is based on the fact that, under the complexity assumptions stipulated, the U-statistics $\widehat{\Theta}_n(\mathcal{P}, \mathbf{P})$ converge to their expectations $\Theta(\mathcal{P}, \mathbf{P})$ uniformly over Π. Naturally, rate bounds could have also been established using similar arguments, following in the footsteps of Section 4.

The following result shows that Rand aggregation preserves consistency under the fast rate condition introduced in Section 5.2.

Proposition 2 (*Rand Median and Consistency*)**.** *Let Π be a collection of partitions of $\mathcal{X}$, which fulfills conditions* (i)–(iii) *in* ***FR****(α). Suppose that there exists $q \in]1, +\infty]$ such that $(\mathbb{E}[D^q(X, X')])^{1/q} < +\infty$. Let $M \geq 1$ and consider M sequences of partitions $\mathbf{P}_N = \{\mathcal{P}_N^{(1)}, \ldots, \mathcal{P}_N^{(M)}\}$ in Π, indexed by $N \geq 1$. Suppose that all these sequences are asymptotically optimal in the clustering risk sense: $\forall m \in \{1, \ldots, M\}$,*

$$W\left(\mathcal{P}_N^{(m)}\right) \to \inf_{\mathcal{P} \in \Pi} W(\mathcal{P}), \quad \text{as } N \to +\infty.$$

Assume that there exists a sequence of $(\overline{\mathcal{P}}_N)_{N \geq 1}$ such that, for all $N \geq 1$, $\overline{\mathcal{P}}_N$ is a Rand consensus of $\mathbf{P}_N$ with respect to Π. Then, the median $\overline{\mathcal{P}}_N$ is asymptotically optimal:

$$W\left(\overline{\mathcal{P}}_N\right) \to \inf_{\mathcal{P} \in \Pi} W(\mathcal{P}), \quad \text{as } N \to +\infty.$$

We point out that one could relax the assumption that the median is taken over the whole set Π at the expense of an additional bias term, asymptotically vanishing. By examining the proof, one may easily see that the result remains true, when assuming that the consensus is taken over a collection $\Pi_N \subset \Pi$, of clustering risk minimizer $\mathcal{P}_N^*$ such that $W(\mathcal{P}_N^*) \to \inf_{\mathcal{P} \in \Pi} W(\mathcal{P})$ as $N \to +\infty$.

Propositions 1 and 2, when combined, show that consistency preservation extends to empirical aggregation of consistent clustering rules, as stated in the following result.

Corollary 1 (*Rand Median and Consistency (Bis)*)**.** *Suppose that Proposition 2's assumptions are satisfied and that the collection of sets $\{\mathcal{C} : \mathcal{C} \in \mathcal{P}, \mathcal{P} \in \Pi\}$ is of finite* VC *dimension. Assume also that an i.i.d. sample $X_1, \ldots, X_n$, with $n \geq 1$, is available, on which the computation of an empirical Rand median $\widehat{\overline{\mathcal{P}}}_N$ of the collection $\mathcal{P}_N$ with respect to the class Π is based. Then, as N and n tend to infinity, we almost-surely have:*

$$W\left(\widehat{\overline{\mathcal{P}}}_N\right) \to \inf_{\mathcal{P} \in \Pi} W(\mathcal{P}).$$

8. Conclusion

Whereas, until now, the theoretical analysis of clustering performance was mainly limited to the K-means situation (but not only, *cf* [13] for instance), this paper establishes bounds for the success of empirical clustering risk minimization in a general "pairwise dissimilarity" framework, relying on the theory of U-processes. The excess of risk of empirical minimizers of the clustering risk is proved to be of the order $O_{\mathbb{P}}(n^{-1/2})$ under mild assumptions on the complexity of the cells forming the partition candidates. It is also shown how to refine slightly this upper bound through a linearization technique and the use of recent inequalities for degenerate U-processes. Under additional assumptions, we also used the same method to establish faster rates of convergence. To the best of our knowledge, the present analysis is the first to state results of this nature. As regards complexity regularization, while focus is here on the choice of the number of clusters, the argument used in this paper also paves the way for investigating more general model selection issues, including choices related to the geometry/complexity of the cells of the partition considered. Finally, we stated preliminary statistical results for empirical clustering aggregation, when consensus is defined in a pairwise manner by means of the Rand distance.

Appendix. Technical proofs

Proof of Theorem 1. Following in the footsteps of Theorem 1's proof in [6], the argument is based on the choice of a finite family $\mathcal{L}_n \subset \mathcal{L}$ of distributions on $[0, 1]^d$ that contains some "bad" distribution, making the selection of the partition that minimizes the clustering risk difficult, and bound by below the supremum by the average behavior over the class $\mathcal{L}_n$.

$$\sup_{\mu \in \mathcal{L}} \mathbb{E}_\mu[W_\mu(\mathcal{P}_n) - \inf_{\mathcal{P} \in \Pi_K} W_\mu(\mathcal{P})] \geq \frac{1}{\#\mathcal{L}_n} \sum_{\mu \in \mathcal{L}_n} \mathbb{E}_\mu[W_\mu(\mathcal{P}_n) - \inf_{\mathcal{P} \in \Pi_K} W_\mu(\mathcal{P})].$$

For simplicity, assume that K is divisible by 3, $K = 3m/2$ say, where $m > 0$ is an integer divisible by 2. The collection $\mathcal{L}_n$ is made of distributions supported by $2m$ fixed points $\{x_i, x_i + w : 1 \leq i \leq m\}$ where $w = (\Delta, 0, 0, \ldots, 0) \in \mathbb{R}^d$ with $\Delta > 0$. It is assumed that for all $i \neq j$, $D(x_i, x_j) = ||x_i - x_j||_p \geq 2\Delta$. The values of the parameters m and $\Delta \geq 1$ are chosen

so that $\Delta = c/\sqrt{m}$ for a constant c suitably chosen, in order to guarantee that one may build a set $\{x_i, x_i + w : \ 1 \leq i \leq m\}$ of points spaced this way in $[0, 1]^d$. Let $\delta \in (0, 1/2)$ and consider the family $\mathcal{L}_n$ of probability distributions μ such that

$$\mu(\{x_i\}) = \mu(\{x_i + w\})$$

for all $i \in \{1, \ldots, m\}$ and which assign mass $(1 - \delta)/(2m)$ to $m/2$ points x_i and mass $(1 + \delta)/(2m)$ to the other $m/2$ points x_j. Precisely, set $\mathcal{L}_n = \{\mu_\gamma : \gamma \in \Gamma\}$ with $\Gamma = \{\gamma \in \{-1, +1\}^m : \ \sum_{j=1}^m \gamma_j = 0\}$, where $\mu_\gamma(x_i) = (1 + \gamma_i\delta)/(2m)$ for $1 \leq i \leq m$. We thus have $\#\mathcal{L}_n = \#\Gamma = m!/((m/2)!)^2$. With no restriction, one may restrict our attention to partitions of the support of the distributions in $\mathcal{L}_n$. Consider the collection $\{\mathcal{P}_\alpha : \ \alpha \in \Gamma\}$ of partitions of $\mathcal{X} = \{x_i, x_i + w : \ 1 \leq i \leq m\}$ such that, for all $\alpha \in \Gamma$, the singletons $\{x_i\}$ and $\{x_i + w\}$ correspond to cells of $\mathcal{P}_\alpha$ when $\alpha_i = +1$, whereas $\{x_i, x_i + w\}$ is a cell of $\mathcal{P}_\alpha$ when $\alpha_i = -1$ (such a partition has then $m + m/2 = K$ cells). We have the following intermediary result.

Lemma 4. *For any partition $\mathcal{P}$ of $\mathcal{X}$ with K cells, there exists $\alpha \in \Gamma$ such that: $\forall \mu \in \mathcal{L}_n$,*

$$W_\mu(\mathcal{P}_\alpha) \leq W_\mu(\mathcal{P}).$$

In addition, for any partition $\mathcal{P}$ of $\mathcal{X}$ with K cells, we have:

$$\forall \gamma \in \Gamma, \quad W_{\mu_\gamma}(\mathcal{P}_\gamma) \leq W_{\mu_\gamma}(\mathcal{P}).$$

Proof. The proof is quite similar to that of steps 3 and 4 in the argument of Theorem 1's proof in [6] and simply relies on the fact that, in our context, we have: $\forall i \neq j$,

$$\min\left\{D(x_i, x_j + w)\mu_\gamma(\{x_j + w\}), D(x_i, x_j)\mu_\gamma(\{x_j\})\right\}\mu_\gamma(\{x_i\}) \geq \Delta\left(\frac{1 - \delta}{2m}\right)^2,$$

which is always larger than $D(x_i, x_i + w)\mu_\gamma(\{x_i\})\mu_\gamma(\{x_i + w\}) = \Delta((1 - \delta)/(2m))^2$ when $\{x_i, x_i + w\}$ forms a cell of $\mathcal{P}_\gamma$. Details are left to the reader. □

Denote by $\mathbf{P}_n$ the collection of empirically designed partitions taking their values in the set $\{\mathcal{P}_\alpha : \ \alpha \in \Gamma\}$ and by U a random variable independent from the X_i's and uniformly distributed on Γ. Consider $\mathcal{P}_{\widehat{\alpha}^*}$ the empirically designed partition defined as follows. For $1 \leq i \leq m$, let $N_i = \sum_{j=1}^n \mathbb{I}\{X_j \in \{x_i, x_i + w\}\}$ and sort the indexes i by increasing order of magnitude of the headcounts: $N_{\sigma(1)} \leq \cdots N_{\sigma(m)}$ with $\sigma \in \mathfrak{S}_m$. For $i \in \{1, \ldots, m/2\}$, we set $\widehat{\alpha}^*_{\sigma(i)} = -1$ ($\{x_{\sigma(i)}, x_{\sigma(i)} + w\}$ is a cell of $\mathcal{P}_{\widehat{\alpha}^*}$), while, for $i \in \{m/2 + 1, \ldots, m\}$, we set $\widehat{\alpha}^*_{\sigma(i)} = +1$ (both $\{x_{\sigma(i)}\}$ and $\{x_{\sigma(i)} + w\}$ are cells of $\mathcal{P}_{\widehat{\alpha}^*}$). Notice that, when the observations $X_1, \ldots, X_n$ are drawn from μ_γ, the random vector $(N_1, \ldots, N_m)$ is distributed as a multinomial vector with parameters $(n;\ q_1, \ldots, q_m)$, where $q_i = (1 + \gamma_i\delta)/\sum_{j=1}^m (1 + \gamma_j\delta)$.

Lemma 5. *We have:*

$$\begin{aligned}\min_{\mathcal{P}_n \in \mathbf{P}_n} \frac{1}{\#\Gamma}\sum_{\gamma \in \Gamma}\left\{\mathbb{E}_{\mu_\gamma}\left[W_{\mu_\gamma}(\mathcal{P}_n)\right] - W_{\mu_\gamma}(\mathcal{P}_\gamma)\right\} &= \min_{\mathcal{P}_n \in \mathbf{P}_n} \mathbb{E}_U\left[\mathbb{E}_{\mu_U}\left[W_{\mu_U}(\mathcal{P}_n)\right] - W_{\mu_U}(\mathcal{P}_U)\right] \\ &\geq \mathbb{E}_U\left[\mathbb{E}_{\mu_U}\left[W_{\mu_U}(\mathcal{P}_{\widehat{\alpha}^*})\right] - W_{\mu_U}(\mathcal{P}_U)\right].\end{aligned}$$

Proof. Observe that, for all $(\gamma, \alpha) \in \Gamma^2$, we have:

$$\begin{aligned}W_{\mu_\gamma}(\mathcal{P}_\alpha) - W_{\mu_\gamma}(\mathcal{P}_\gamma) &= \frac{\Delta}{2m^2}\sum_{i=1}^m \frac{1 - \alpha_i}{2}(1 + \gamma_i\delta)^2 - (1 - \delta)^2\frac{\Delta}{2m^2} \\ &= \frac{\Delta\delta}{2m^2}\sum_{i=1}^m \gamma_i(1 - \alpha_i) + \frac{\Delta}{2m^2}\left\{\frac{m}{2}(1 + \delta^2) - (1 - \delta)^2\right\}.\end{aligned}$$

Noticing that it suffices to focus on empirically designed partitions $\mathcal{P}_{\widehat{\alpha}}$ which depend on the empirical counts $N_1, \ldots, N_m$, consider $\widehat{\alpha} = \widehat{\alpha}(N_1, \ldots, N_m)$, taking its values in Γ. Observe that one may write

$$\mathbb{E}\left[W_{\mu_\gamma}(\mathcal{P}_{\widehat{\alpha}}) - W_{\mu_\gamma}(\mathcal{P}_{\widehat{\alpha}^*})\right] = \frac{\Delta\delta}{2m^2}\sum_{i=1}^m \gamma_i(\alpha_i^*(n_1, \ldots, n_m) - \alpha_i(n_1, \ldots, n_m))\mathbb{P}_{\mu_\gamma}\{(N_1, \ldots, N_m) = (n_1, \ldots, n_m)\}.$$

By virtue of Step 7 in Theorem 1's proof in [6], the quantity given by $\sum_{i=1}^m \gamma_i(\alpha_i^* - \alpha_i)\mathbb{P}_{\mu_\gamma}\{(N_1, \ldots, N_m) = (n_1, \ldots, n_m)\}$ is nonnegative for any $\gamma \in \Gamma$, which establishes the lemma. □

By symmetry, we have:

$$\mathbb{E}_U\left[\mathbb{E}_{\mu_U}\left[W_{\mu_U}(\mathcal{P}_{\widehat{\alpha}^*})\right] - W_{\mu_U}(\mathcal{P}_{(U)})\right] = \mathbb{E}_{\mu_\gamma}\left[W_{\mu_\gamma}(\mathcal{P}_{\widehat{\alpha}^*})\right] - W_{\mu_\gamma}(\mathcal{P}_\gamma),$$

for any fixed $\gamma \in \Gamma$. Now, denote by p_j the probability that the empirically optimal partition makes $j \in \{1, \ldots, m/2\}$ mistakes. Taking $\delta = \sqrt{m/n}$, we have:

$$\mathbb{E}_{\mu_\gamma}\left[W_{\mu_\gamma}(\mathcal{P}_{\widehat{\alpha}^*})\right] - W_{\mu_\gamma}(\mathcal{P}_\gamma) = 2\Delta\frac{\delta}{m^2}\sum_{j=1}^{m/2} j p_j$$
$$\geq \frac{\Delta\delta}{m}\frac{\Phi(-2)^4}{128} = \frac{c}{(2K/3)^{3/2}}\frac{\Phi(-2)^4}{128}\frac{1}{\sqrt{n}},$$

using the lower bound for $\sum_{j=1}^{m/2} j p_j$ established in Theorem 1's proof in [6] (see Steps 10–11 therein). This permits to finish the proof.

Proof of Theorem 2. We may classically write:

$$\widehat{W}(\widehat{\mathcal{P}}_n) - W_K^* \leq 2\sup_{\mathcal{P}\in\Pi_K}|\widehat{W}_n(\mathcal{P}) - W(\mathcal{P})| + \inf_{\mathcal{P}\in\Pi_K} W(\mathcal{P}) - W_K^*$$
$$\leq 2K\sup_{\mathcal{C}\in\mathcal{P},\mathcal{P}\in\Pi_K}|U_n(\mathcal{C}) - u(\mathcal{C})| + \inf_{\mathcal{P}\in\Pi_K} W(\mathcal{P}) - W_K^*, \tag{25}$$

where $U_n(\mathcal{C})$ denotes the U-statistic with kernel given by $\mathcal{H}_{\mathcal{C}}(x, x') = D(x, x')\cdot\mathbb{I}\{(x, x') \in \mathcal{C}^2\}$ and based on the sample $X_1, \ldots, X_n$ and $u(\mathcal{C})$ its expectation. Therefore, mimicking the argument of Corollary 3 in [15], based on the so-termed *first Hoeffding's representation* of U-statistics (see Lemma A.1 in [15]), we may straightforwardly derive the lemma below.

Lemma 6 (*Uniform Deviations*). *Suppose that Theorem 2's assumptions are fulfilled. Let $\delta > 0$. With probability at least $1 - \delta$, we have: $\forall n \geq 2$,*

$$\sup_{\mathcal{C}\in\mathcal{P},\mathcal{P}\in\Pi_K}|U_n(\mathcal{C}) - u(\mathcal{C})| \leq 2\mathbb{E}[\mathcal{A}_{K,n}] + B\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{26}$$

Proof. The argument follows in the footsteps of Corollary 3's proof in [15]. It is based on the so-termed *first Hoeffding's representation* of U-statistics (9), which provides an immediate control of the moment generating function of the supremum $\sup_{\mathcal{C}}|U_n(\mathcal{C}) - u(\mathcal{C})|$ by that of the norm of an empirical process, namely $\sup_{\mathcal{C}}|A_n(\mathcal{C}) - u(\mathcal{C})|$, where, for all $\mathcal{C} \in \mathcal{P}$ and $\mathcal{P} \in \Pi_K$:

$$A_n(\mathcal{C}) = \frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} D(X_i, X_{i+\lfloor n/2\rfloor})\cdot\mathbb{I}\{(X_i, X_{i+\lfloor n/2\rfloor}) \in \mathcal{C}^2\}.$$

Lemma 7 (*See Lemma A.1 in [15]*). *Let $\Psi : \mathbb{R} \to \mathbb{R}$ be convex and nondecreasing. We have:*

$$\mathbb{E}\left[\exp\left(\lambda\cdot\sup_{\mathcal{C}}|U_n(\mathcal{C}) - u(\mathcal{C})|\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda\cdot\sup_{\mathcal{C}}|A_n(\mathcal{C}) - u(\mathcal{C})|\right)\right]. \tag{27}$$

Now, using standard symmetrization and randomization tricks, one obtains that: $\forall\lambda > 0$,

$$\mathbb{E}\left[\exp\left(\lambda\cdot\sup_{\mathcal{C}}|A_n(\mathcal{C}) - u(\mathcal{C})|\right)\right] \leq \mathbb{E}\left[\exp\left(2\lambda\cdot\mathcal{A}_{K,n}\right)\right]. \tag{28}$$

Observing that the value of $\mathcal{A}_{K,n}$ cannot change by more than $2B/n$ when one of the $(\epsilon_i, X_i, X_{i+\lfloor n/2\rfloor})'s$ is changed, while the others are kept fixed, the standard bounded differences inequality argument applies and yields:

$$\mathbb{E}\left[\exp\left(2\lambda\cdot\mathcal{A}_{K,n}\right)\right] \leq \exp\left(2\lambda\cdot\mathbb{E}[\mathcal{A}_{K,n}] + \frac{\lambda^2 B^2}{2n}\right). \tag{29}$$

Next, Markov's inequality with $\lambda = (t - 2\mathbb{E}[\mathcal{A}_{K,n}])/B^2$ gives: $\mathbb{P}\{\sup_{\mathcal{C}}|A_n(\mathcal{C}) - u(\mathcal{C})| > t\} \leq \exp(-n(t - 2\mathbb{E}[\mathcal{A}_{K,n}])^2/(2B^2))$. The desired result is then immediate. □

The rate bound is finally established by combining bounds (25) and (26).

Proof of Theorem 3 (*Sketch of*). The theorem can be proved by using the decomposition (10), applying the argument above in order to control $\sup_{\mathcal{P}}|L_n(\mathcal{P})|$ and the lemma below to handle the degenerate part. The latter is based on a recent moment inequality for degenerate U-processes, proved in [15]. Technical details are left to the reader.

Lemma 8 (*See Theorem 11 in [15]*). *Suppose that Theorem 3's assumptions are fulfilled. There exists a universal constant $C < \infty$ such that for all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$: $\forall n \geq 2$,*

$$\sup_{\mathcal{P} \in \Pi_K} |M_n(\mathcal{P})| \leq K\kappa(n, \delta).$$

Proof of Theorem 4. By virtue of Bernstein's exponential probability inequality combined with the union bound, for all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:

$$\forall \mathcal{P} \in \Pi, \quad 0 \leq \mathcal{L}_n(\mathcal{P}) + \sqrt{\frac{2\,\mathrm{Var}(\mathcal{H}_{\mathcal{P}}^{(1)}(X))\log(N/\delta)}{n}} + \frac{4\log(N/\delta)}{3n}.$$

Combining Lemmas 1 and 2 with the union bound again and the fact that $\Lambda_n(\widehat{P}_n^*) \leq 0$, we obtain that, with probability at least $1 - \delta$:

$$W(\widehat{\mathcal{P}}_n^*) - W^* \leq 2\sqrt{\frac{2B^2\left(W(\widehat{\mathcal{P}}_n^*) - W^*\right)^{\alpha}\log(2N/\delta)}{n}} + \frac{4\log(2N/\delta)}{3n} + c\frac{\log(2N/\delta)}{n}.$$

The desired result is then obtained by solving this inequality for $W(\widehat{\mathcal{P}}_n^*) - W^*$.

Proof of Theorem 5. The proof mimics the argument of Theorem 8.1 in [10]. We thus obtain that: $\forall K \geq 1$,

$$\mathbb{E}\left[W(\widehat{\mathcal{P}}_{K,n})\right] - W^* \leq \mathbb{E}\left[W(\widehat{\mathcal{P}}_{K,n})\right] - W^* + pen(K, n) + \sum_{k \geq 1}\mathbb{E}\left[\left(\sup_{\mathcal{P} \in \Pi_k}\{W(\mathcal{P}) - \widehat{W}_n(\mathcal{P})\} - pen(n, k)\right)_+\right].$$

Reproducing the argument of Theorem 2's proof, one may easily show that: $\forall k \geq 1$,

$$\mathbb{E}\left[\sup_{\mathcal{P} \in \Pi_k}\{W(\mathcal{P}) - \widehat{W}_n(\mathcal{P})\}\right] \leq 2k\mathbb{E}[\mathcal{A}_{k,n}].$$

Thus, for all $k \geq 1$, the quantity $\mathbb{P}\{\sup_{\mathcal{P} \in \Pi_k}\{W(\mathcal{P}) - \widehat{W}_n(\mathcal{P})\} \geq pen(n, k) + 2\delta\}$ is bounded by

$$\mathbb{P}\left\{\sup_{\mathcal{P} \in \Pi_k}\{W(\mathcal{P}) - \widehat{W}_n(\mathcal{P})\} \geq \mathbb{E}\left[\sup_{\mathcal{P} \in \Pi_k}\{W(\mathcal{P}) - \widehat{W}_n(\mathcal{P})\}\right] + \sqrt{(2B\log k)/n} + \delta\right\}$$
$$+ \mathbb{P}\left\{3k\mathbb{E}_{\epsilon}[\mathcal{A}_{k,n}] \leq 2k\mathbb{E}[\mathcal{A}_{k,n}] - \frac{27Bk\log k}{n} - \delta\right\}.$$

By virtue of the bounded differences inequality (jumps being bounded by $2B/n$), the first term is bounded by $\exp(-n\delta^2/(2B^2))/k^2$, while the second term is bounded by, $\exp(-n\delta/(9Bk))/k^3$ as shown by Lemma 8.2 in [10] (see the third inequality therein). Integrating over δ, one obtains:

$$\mathbb{E}\left[\left(\sup_{\mathcal{P} \in \Pi_k}\{W(\mathcal{P}) - \widehat{W}_n(\mathcal{P})\} - pen(n, k)\right)_+\right] \leq (2B\sqrt{2/n} + 18B/n)/k^2.$$

Summing next the bounds thus obtained over k leads to the oracle inequality stated in the theorem.

Proof of Proposition 1. Observe that

$$0 \leq \Theta(\widehat{\overline{\mathcal{P}}}_n, \mathbf{P}) - \Theta(\overline{\mathcal{P}}_n, \mathbf{P}) \leq 2\sup_{\mathcal{P} \in \Pi}\left|\widehat{\Theta}_n(\mathcal{P}, \mathbf{P}) - \Theta(\mathcal{P}, \mathbf{P})\right|.$$

Equipped with the notations introduced in Section 7, for any $\mathcal{P} \in \Pi$, the kernel of the U-statistic $\widehat{\Theta}_n(\mathcal{P})$ is given by $\sum_{m=1}^{M}\mathcal{K}_{\mathcal{P},\mathcal{P}^{(m)}}(x, x')$. Noticing that, under the assumption that the collection of cells forming the partitions in Π (respectively, in $\mathbf{P}$) is of finite VC dimension, this collection of kernels forms a VC major class of functions (see [20]) of finite VC dimension, the strong Law of Large Numbers for U-processes stated in Corollary 5.2.3 in [19] shows that the term on the right hand side of the bound above vanishes almost-surely, as $n \to +\infty$. In addition, the CLT for U-processes given in Theorem 5.3.7 by [19] proves that this convergence holds at the rate $O_{\mathbb{P}}(n^{-1/2})$.

Proof of Proposition 2. Using the triangular inequality, we get

$$d_R\left(\overline{\mathcal{P}}_N, \mathcal{P}^*\right) \leq \frac{1}{M}\sum_{m=1}^{M}\left\{d_R(\mathcal{P}_N^{(m)}, \mathcal{P}^*) + d_R(\overline{P}_N, \mathcal{P}^{(m)})\right\}$$
$$\leq \frac{2}{M}\sum_{m=1}^{M} d_R(\mathcal{P}_N^{(m)}, \mathcal{P}^*),$$

since we assumed $\mathcal{P}^* \in \Pi$. As condition (iii) in **FR**(α) entails that, for all $m \in \{1, \ldots, M\}$, $d_R(\mathcal{P}_N^{(m)}, \mathcal{P}^*) \to 0$ as $N \to +\infty$, this bound shows that $d_R\left(\overline{\mathcal{P}}_N, \mathcal{P}^*\right) \to 0$ as $N \to +\infty$. Combined with Lemma 3, this establishes the desired convergence.

Proof of Corollary 1. Re-using the bounds involved in the two previous proofs, we obtain that:

$$\begin{aligned} d_R(\widehat{\overline{\mathcal{P}}}_n, \mathcal{P}^*) &\leq \frac{2}{M} \sup_{\mathcal{P} \in \Pi} |\Theta(\mathcal{P}, \mathbf{P}) - \widehat{\Theta}_n(\mathcal{P}, \mathbf{P})| + d_R(\overline{\mathcal{P}}_N, \mathcal{P}^*) \\ &\leq 2 \sup_{(\mathcal{P}, \mathcal{P}') \in \Pi^2} |d_R(\mathcal{P}, \mathcal{P}') - \widehat{d}_R(\mathcal{P}, \mathcal{P}')| + d_R(\overline{\mathcal{P}}_N, \mathcal{P}^*). \end{aligned}$$

The first term on the right hand side of the bound above vanishes as $n \to +\infty$ with probability one, by virtue of the same SLLN argument as that used for proving Proposition 1. The argument of Proposition 2's proof shows next that the second term converges to zero as $N \to +\infty$. Lemma 3 permits then to finish the proof.

References

[1] N. Ailon, M. Charikar, A. Newman, Aggregating inconsistent information: Ranking and clustering, Journal of the ACM 55 (5) (2008) 23:1–23:27.
[2] A. Antos, L. Györfi, A. György, Individual convergence rates in empirical vector quantizer design, IEEE Transaction on Information Theory 51 (11) (2005) 4013–4023.
[3] M.A. Arcones, E. Giné, Limit theorems for U-processes, The Annals of Probability 21 (3) (1993) 1494–1542.
[4] J.-P. Barthélemy, B. Leclerc, The median procedure for partitions, in: DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 19, 1995, pp. 3–34.
[5] J.P. Barthélémy, B. Montjardet, The median procedure in cluster analysis and social choice theory, Mathematical Social Sciences 1 (1981) 235–267.
[6] P.L. Bartlett, T. Linder, G. Lugosi, The minimax distorsion redundancy in empirical quantizer design, IEEE Transaction on Information Theory 44 (5) (1998) 1802–1813.
[7] S. Ben-David, A framework for statistical clustering with a constant time approximation algorithms for k-median clustering, in: Proceedings of COLT'04, in: Lecture Notes in Computer Science, vol. 3120, 2004, pp. 415–426.
[8] G. Biau, L. Bleakley, Statistical inference on graphs, Statistics & Decisions 24 (2006) 209–232.
[9] G. Biau, L. Devroye, G. Lugosi, On the performance of clustering in hilbert space, IEEE Transaction on Information Theory 54 (2) (2008) 781–790.
[10] S. Boucheron, O. Bousquet, G. Lugosi, Theory of classification: a survey of some recent advances, ESAIM: Probability and Statistics 9 (2005) 323–375.
[11] L. Breiman, Bagging predictors, Machine Learning 26 (1996) 123–140.
[12] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[13] S. Bubeck, U. von Luxburg, Nearest neighbor clustering: a baseline method for consistent clustering with arbitrary objective functions, Journal of Machine Learning Research 10 (2009) 657–698.
[14] B. Clarke, E. Fokoué, H. Zhang, Principles and Theory for Data-Mining and Machine-Learning, Springer, 2009.
[15] S. Clémençon, G. Lugosi, N. Vayatis, Ranking and empirical risk minimization of U-statistics, The Annals of Statistics 36 (2) (2008) 844–874.
[16] S. Clémençon, N. Vayatis, Ranking the best instances, Journal of Machine Learning Research 8 (2007) 2671–2699.
[17] S. Clémençon, N. Vayatis, Overlaying classifiers: a practical approach to optimal scoring, Constructive Approximation 32 (2010) 619–648.
[18] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, 1996.
[19] V. de la Pena, E. Giné, Decoupling: From Dependence to Independence, Springer, 1999.
[20] R.M. Dudley, Uniform Central Limit Theorems, Cambridge University Press, 1999.
[21] D.K. Fuk, S.V. Nagaev, Probability ineaqualities for sums of independent random variables, Probability Theory and its Applications 16 (4) (1971) 643–660.
[22] J.A. Hartigan, Asymptotic distributions for clustering criteria, The Annals of Statistics 6 (1978) 117–131.
[23] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer, 2009, pp. 520–528.
[24] W. Hoeffding, A class of statistics with asymptotically normal distribution, Annals of Mathematical Statistics 19 (1948) 293–325.
[25] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1) (1985) 193–218.
[26] O. Hudry, NP-Harness of the computation of a median equivalence relation in classification, Mathematics and Social Sciences 197 (1) (2012) 83–97.
[27] V. Koltchinskii, Local Rademacher complexities and oracle inequalities in risk minimization, The Annals of Statistics 34 (2006) 2593–2706. with discussion.
[28] S. Kutin, P. Niyogi, Almost-everywhere algorithmic stability and generalization error, in: Proceedings of the of the 18th Conference in Uncertainty in Artificial Intelligence, 2002.
[29] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1999.
[30] P. Massart, Concentration Inequalities and Model Selection, in: Lecture Notes in Mathematics/cole d't de Probabilits de Saint-Flour, Springer, 2003.
[31] P. Massart, E. Nédélec, Risk bounds for statistical learning, Annals of Statistics 34 (5) (2006).
[32] R. Peck, L. Fisher, J. van Ness, Bootstrap confidence intervals for the number of clusters in cluster analysis, Journal of the American Statistical Association 84 (1989) 184–191.
[33] D. Pollard, Strong consistency of k-means clustering, The Annals of Statistics 9 (1981) 135–140.
[34] D. Pollard, A central limit theorem for k-means clustering, The Annals of Probability 10 (1982) 919–926.
[35] W. Rand, Objective criteria for the evaluation of clustering methods, J. Amer. Stat. Assoc. 66 (336) (1971) 846–850.
[36] R.J. Serfling, Approximation Theorems of Mmathematical Statistics, Wiley, 1980.
[37] O. Shamir, N. Tishby, On the reliability of clustering stability in the large sample regime, in: Advances in Neural Information Processing Systems, vol. 21, 2009.
[38] O. Shamir, N. Tishby, Model selection and stability in k-means clustering, in: Proceedings of the 21rst Annual Conference on Learning Theory, 2008.
[39] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society 63 (2) (2001) 411–423.
[40] A. Tsybakov, Introduction à l'estimation non-paramétrique, in: Mathématiques et Applications, Springer, 2004.
[41] A. Tsybakov, Optimal aggregation of classifiers in statistical learning, Annals of Statistics 32 (1) (2004) 135–166.
[42] A. van der Vaart, Asymptotic Statistics, Cambridge University Press, 1998.
[43] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance, JMLR 11 (2010) 2837–2854.
[44] U. von Luxburg, Clustering stability: an overview, Foundations and Trends in Machine Learning 2 (3) (2009) 235–274.
[45] U. von Luxburg, S. Ben-David, Towards a statistical theory of clustering, in: Pascal Workshop on Statistics and Optimization of Clustering, 2005.
[46] U. von Luxburg, S. Ben-David, A sober look at clustering stability, in: Proceedings of the 19th Conference on Learning Theory, 2006.
[47] U. von Luxburg, S. Ben-David, Relating clustering stability to properties of cluster boundaries, in: Proceedings of the 21th Conference on Learning Theory, 2008.
[48] D.M. Witten, R. Tibshirani, A framework for feature selection in clustering, Journal of the American Statistical Association 105 (490) (2010) 713–726.